Organisation for Economic Co-operation and Development

ENV/JM/MONO(2019)18

Unclassified

English - Or. English

28 August 2019

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

VALIDATION REPORT OF THE XENOPUS ELEUTHEROEMBRYONIC THYROID SIGNALING ASSAY (XETA) FOR THE DETECTION OF THYROID ACTIVE SUBSTANCES

Series on Testing and Assessment
No. 302

JT03450433

SERIES ON TESTING AND ASSESSMENT
NO. 302

VALIDATION REPORT OF THE XENOPUS ELEUTHEROEMBRYONIC THYROID SIGNALING ASSAY (XETA) FOR THE DETECTION OF THYROID ACTIVE SUBSTANCES

IOMC
INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS
A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate
ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT
Paris 2019

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 35 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in eleven different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides**; **Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents;** and **Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

This document describes the design and results of two phases of the validation effort for the Xenopus Eleutheroembryonic Thyroid signaling Assay (XETA). This method was developed for the detection of thyroid active substances. It is performed in 6-well plate format and can serve as a quick screen for potential thyroid disrupting substances. The purpose of phase I of the validation was to determine whether the standard operating procedure (SOP) can be used across international laboratories. Reliability and reproducibility of the SOP was established for a series of chemicals tested across multiple laboratories. Two of the laboratories were naïve to the protocol, while the third was an expert laboratory. The purpose of the phase II validation was to generate a larger set of data to confirm the conclusions from phase I regarding the transferability of the assay and to set a definitive statistical protocol for the analysis of the results.

The XETA assay has been validated through an international effort via the OECD. The OECD has been working with member countries on the validation and harmonization of testing methods for the detection of chemicals that interfere with the estrogen, androgen and thyroid pathways.

The project to develop this Guidance Document was led by France. The Validation Management Group for Ecotoxicity (VMG-Eco) assisted the validation exercise to develop the XETA by evaluating the validation proposals as well as the validation exercise results.

The Working Group of the National Coordinators of the Test Guidelines Programme endorsed the validation report at its 31st meeting in April 2019.

Table of contents

FIGURES

TABLES

ABREVIATIONS AND DEFINITIONS

GFP: Green fluorescent protein
LC50: Median lethal concentration is the concentration of a test chemical that is estimated to be lethal to 50% of the test organisms within the test duration
LOEC: The lowest observed effect concentration is the lowest tested concentration at which the test chemical is observed to have a statistically significant effect
MS-222: tricaine methanesulfonate
MTC: Maximum tolerated concentration
NOEC: The no observed effect concentration is the tested concentration immediately below the LOEC.
PTU: Propylthiouracil
SEM: Standard error to the mean
SMILES: Simplified molecular input line entry specification.
Runs: the number of experiment performed for each chemical. Each run utilizes different spawn and test solutions independently prepared
Spiked mode: Part of the XETA run in the presence of 3.25µg/l of T3
T3: Triiodothyronine
T4: thyroxine
TR: thyroid receptor
TH: thyroid hormones
THbZIP: Thyroid hormone beta zip transcription factor
Unspiked mode: Part of the XETA run in the absence of T3
UVCB: Substances of unknown or variable composition, complex reaction products or biological materials.

ACKNOWLEDGEMENTS

This work is the collaborative effort of six laboratories which generously performed the experiments outlined. John Green, DuPont, performed the statistical analysis and the power simulations.

The following laboratories and their staff performed the testing for the XETA phase I validation:

- Dr Taisen Iguchi, the Department of Bio-Environmental Science, Center for Integrative Bioscience, Okazaki National Research Institutes, Okazaki, Japan: Dr Ikumi Hirakawa and Dr Yukiko Ogino performed the experiments.
- Dr Daniel Buchholz, McMicken College of Arts and Sciences, Department of Biological Sciences, University of Cincinnati, Cincinnati, USA: Allyson J. Cameron and Alison Brittain (master students) performed the experiments.
- Laboratoire WatchFrog, Evry, France: Anthony Sébillot (research engineer) performed the experiments. Dr David Du Pasquier designed and coordinated the phase 1 validation experiments.

The following laboratories and their staff performed the testing for the XETA phase II validation:

- Dr Taisen Iguchi (Okazaki National Research Institutes) and Dr Yuta Onishi (Institute of Environnemental Ecology, IDEA Consultants, Inc.), Japan. Tetsuro Okamura, Yasushi Goto, Maki Sakurai, Jun Yamamoto, Yu Totsuka at IDEA performed the XETA experiments and the chemical analysis.
- Dr Isabel Lopes, Departamento de Biologia & CESAM, Universidade de Aveiro, Portugal. Isabel Lopes, Marta Monteiro (postdoctoral student) and Carla Quintaneiro (postdoctoral student) performed the experiments.
- Dr Jean Francois Mougel, AQUIRIS Laboratory, Brussels, Belgium. Marie Stievenard (research engineer) performed the experiments.
- Laboratoire WatchFrog, Evry, France : Anthony Sébillot (research engineer) and Kilian Lamiral (master student) performed the experiments. Dr David Du Pasquier designed and coordinated the phase II validation experiments.

David Du Pasquier (pasquier@watchfrog.fr), wrote the validation report and the draft test guideline.

1. INTRODUCTION

1.1. Objectives of the Validation Study

The overall objective of the validation exercise for the XETA is to establish the relevance of the assay to detect thyroid activity of compounds acting at different points within the thyroid system. A second aim is to assess the transferability and reproducibility of the assay by comparing results obtained by a variety of laboratories in diverse geographical locations.

The phase I validation focused on demonstrating the relevance of the assay, i.e. its ability to detect compounds that act on the thyroid system. Phase I also showed that the assay protocol is optimized for performance across laboratories.

The phase II validation included a demonstration of the benefits of the efficiency and medium-throughput-nature of this test in screening for potential thyroid active compounds. Phase II examined the ability of the protocol to test a wider range of thyroid active chemicals with different modes of action (expanding the modes that were tested in phase I) and additional negative substances to challenge the protocol and the statistical method of analysis.

1.2. Assay Development/Background

Transferability of the XETA was successfully demonstrated in three different French laboratories (CNRS, Paris; WatchFrog, Evry and Institut Pasteur de Lille, Maxeville) before the OECD validation started.

To date data using the XETA have been ratified in 10 publications (Turque et al. 2005, Fini et al. 2007, 2009, 2017; Castillo et al. 2013; Neal et al 2017; Spirhanzlova et al. 2017 ; Välitalo et al. 2017, Escher et al. 2018 and Leusch et al. 2018).

Fini et al. 2007 demonstrated that the assay detects a series of substances interfering with endogenous TH metabolization or TH action at multiple levels. These substances included antagonists acting at the level of the receptor or the thyroid gland, as well as agonists. Further

data on testing other chemicals are included in Fini et al. 2012 and 2017; Spirhanzlova et al. 2017 and Neal et al. 2017.

Fini et al. 2009 and Fini et al. 2012 show that the *Xenopus laevis (X.laevis)* embryonic model is metabolically competent to conjugate and excrete xenobiotics using enzymatic pathways which are homologous to mammalian pathways.

Castillo et al. 2013, Leusch 2017, Välitalo et al. 2017 and Escher et al. 2018 use the XETA for the assessment of water quality. In a comparative study including available *in vitro* tests Leusch et al. 2017 conclude that the XETA is suitable and sufficiently sensitive to detect thyroid active compounds in environmental water samples.

The following information supported the transferability of this protocol to the participating laboratories for phase I and II testing:

1. *Xenopus laevis* is a much-used model in many industrial, government and academic laboratories. The reasons for this are multiple. Besides being a major model for developmental biologists, toxicologists have long used the *Xenopus laevis* embryo for the FETAX test (Frog Embryonic Toxicology Assay – Xenopus (ASTM E1439-12) and more recently for the AMA and LAGDA tests. Thus, many institutions and Contract Research Organisations are equipped for breeding or maintaining Xenopus as adults or providing stabulation of tadpoles/embryos.
2. Founders, adult *Xenopus laevis* for breeding embryos for on-site testing, were made available to participants in the ring test.
3. Embryos were available for shipping from a breeding/production site to another testing site.

1.3. Test organism

The South African clawed frog, *Xenopus laevis,* is the test species selected for the XETA. *Xenopus laevis* is an especially useful Amphibian model because its early development and metamorphosis have been extensively studied. In addition, amphibians and higher vertebrates share a great amount of genetic homology, allowing the XETA assay to provide information that may be extrapolated to other organisms. This test species is also utilized in OECD Test Guidelines

focused on Amphibians (Amphibian Metamorphosis Assay ; OECD TG 231 and LAGDA TG 241). The welfare of the animal in this assay will be of major concern. References for documents describing the guidance and care of *Xenopus laevis* are included here (Barney 2005; Green 2009). Xenopus is an excellent vertebrate model for biotransformation studies and endocrine disruption studies displaying homologous pathways to mammals (Fini et al. 2012a; 2012b).

Amphibians are well established as a reference model for the biological consequences of thyroid disruption. The following characteristics are particularly well known with respect to the thyroid system of *Xenopus laevis* (Shi 2000):

- Activation of thyroid signalling leads to an acceleration of metamorphosis in the larva, which induces: cell death; resorption of pre-metamorphic tissues (such as the gills and the tail); the growth of limb buds; and death if the metamorphosis is triggered too early during development.
- Inhibition induces a delay in or a complete absence of metamorphosis, and the animals remain in the larval stage but continue to grow (leading to giant tadpoles).

In humans, given the role of thyroid hormones, particularly in the developing foetal brain, heart and metabolic regulation, disruption of the thyroid axis may have similar consequences (Yen, 2001, Demeneix 2014). The fact that thyroid hormone signalling is conserved through vertebrates, with exactly the same hormones in all vertebrates and with high homology in the different components, was the rational for the amphibian metamorphosis test that is already recognized as providing information that is relevant to humans. Clearly the high homology of Xenopus receptors and transport proteins to human counterparts is an argument to be brought forward, as is the parallel metabolism of xenobiotics (Fini et al. 2012).

1.4. Genetic construct

The XETA is performed during eleutheroembryonic stages of development (stage NF45 up to stage NF47) using a transgenic line of *Xenopus laevis* called the THbZIP-GFP line. Transgenic tadpoles are conceived from natural mating of wild type females with males of the THbZIP-GFP homozygous line (founders).

Transgenic tadpoles carry a portion of the *Xenopus laevis THbZIP* promoter cloned upstream of the Green Fluorescent Protein (GFP) coding sequence and flanked by two copies of insulator sequences of the chicken lysozyme gene. The use of insulators helps to achieve a more homogeneous basal level of expression of the transgene, and may protect the transgenes from methylation, thus maintaining transgene expression in descendants (Kirillov et al. 1996).

A schematic of the construct is shown in Figure 1 below.

Figure 1: Genetic Construct of THbZIP.

The gene THbZIP encodes for a transcription factor that contains basic domains and leucine zipper (Wang et al. 1993). The expression of THbZIP is regulated by endogenous and exogenous thyroid hormones (Turque et al. 2005).

The genetic construct contains the portion of the *THbZIP* gene promoter between -246bp and 130bp and includes two TREs (Thyroid Responsive Elements). This promoter controls the expression of a fluorescent reporter protein (GFP). Thus, the level of fluorescence detected is proportional to the level of transcription of the GFP gene. As the transcription of this construct is regulated by the activation / transport of the TH and the subsequent binding of the TH-Thyroid receptor complex to the TREs, chemicals interfering with the thyroid axis can be detected via modulation in the level of GFP expression (Fini et al. 2007; 2009).

2. PURPOSE AND OBJECTIVES

2.1. Purpose of the assay

XETA is an aqueous assay that utilizes free-living *X. laevis* eleutheroembryonic-stage animals in a multi-well format to detect modulation of thyroid receptor signalling by thyroid active chemicals. The assay is transcription-based and uses a *X. laevis* transgenic line harbouring the THbZIP-GFP genetic construct. This transgenic line is commercially available (see 4.6.3. Transgenic line availability). The assay measures the ability of a chemical to activate or inhibit transcription of the genetic construct, whether directly through binding to the thyroid receptor (TR) or modifying the binding of thyroid hormones (TH) to the TR, or indirectly by modifying the amount of TH available to activate the TR and thereby transcription of the THbZIP-GFP construct (see Figure 1 above).

To date the XETA has been shown to detect chemicals acting through various mechanisms of action including TH receptors, agonists (e.g. T4, TRIAC) and antagonists (e.g. NH3 (a pharmacological antagonist of the TRs)), modulators of TH clearance (including UDPGT (UDP-glucuronosyltransferase) modulators (e.g. Phenobarbital) and modulators of TH metabolism (including deiodinase inhibitors (e.g. iopanoic acid) (Fini et al. 2007). In addition, the XETA potentially detects modulators of TH transport via interaction with TH plasma binding proteins and inhibitors of TH transmembrane transporters.

As Xenopus NF45 stage embryos are not synthesising their own TH, inhibitors of TH synthesis are not detected by the XETA. The XETA does not distinguish between the different modes of action but provide information on whether a chemical acts as a global activator or inhibitor of the thyroid signalling pathway in the *X. laevis* eleutheroembryo.

The endpoint measured is fluorescence of tadpoles. When transcription of the genomic construct is activated or inhibited following chemical exposure, tadpoles express more or less GFP and therefore emit more or less fluorescence compared to unexposed tadpoles where fluorescence remains at the basal level.

2.2. Major characteristics of the assay

The XETA was designed as a screening assay to provide information only on the potential of a test chemical to alter the normal functions of

the thyroid system. The XETA provides a rapid (72h exposure time) assay for measuring the response of eleutheroembryonic stage tadpoles to potential thyroid active chemicals.

The assay measures GFP protein fluorescence in the transgenic tadpoles by way of a 96-well plate fluorescence reader or a fluorescent microscope equipped with a specific camera (Annex 1) that transforms the fluorescence signal to a numerical value. Each tadpole expresses a basal fluorescence in its translucent tissues that corresponds to the control value.

If a thyroid active chemical modulates the transcription of the THbZIP-GFP construct, the GFP protein production is either down or up regulated:

- a pro-thyroid compound is revealed by an increase of fluorescence
- an anti-thyroid compound is revealed by either an increase or a decrease, depending on the mode of action.

An internal control, the most biologically active form of thyroid hormone, T3 (Triiodothyronine), establishes the level of activation of the axis for each test series. This percentage increase in fluorescence corresponds to the signal induced by a concentration of T3 (3.25 µg/L) that is equivalent to the plasma T3 concentration during tadpole metamorphosis (Leloup et Buscaglia1977).

2.3. General experimental design

The assay is performed to determine the potential of a test chemical to modulate the thyroid system under sublethal concentrations. For the validation process a five concentrations test design has been used (10 tadpoles per well x 2 wells = 20 tadpoles exposed per concentration). In the test guideline a minimum of three concentrations is recommended allowing to keep the same sensitivity. Tadpoles are used for the XETA between developmental stages NF45 (beginning of the test) and 47 (end of the test). They are not fed during the test as yolk is still present in the intestine from stage NF45 to stage NF47 and is used as the source of energy for the development of the tadpole (Nieuwkoop and Faber 1994).

The test is run in two modes “spiked” and “unspiked” i.e. with and without the addition of T3. In spiked mode all groups are spiked with 3.25 µg/l of T3, a physiological concentration corresponding to the

plasmatic T3 concentration, at the climax of *X. laevis* metamorphosis (Leloup et Buscaglia 1977). Tadpoles at stage NF45 are not synthetizing T3 and only low T3 concentration is detectable in the tadpole body that come from TH accumulated in the egg. Therefore, this spiking is necessary to detect chemicals acting on TH distribution, metabolization, degradation and TR antagonists.

The control groups include:

- -Twenty tadpoles exposed to test medium only ("negative controls").
- -Twenty tadpoles exposed to T3 at one concentration: 3.25 µg/L ("T3 controls"). This control serves as a positive control for the part of the test without T3 added and a negative control for the part of the test with T3.
- -Twenty tadpoles exposed to T3 at 3.25 µg/L and T4 at a high concentration, 10 mg/l ("T3 +T4 controls"). This control is included in each assay to confirm that the fluorescence obtained with T3 is not saturated. It serves also as a positive control for the part of the test with T3 added.

After 72 hours of exposure, the tadpoles are anesthetized and placed individually on their dorsal sides into wells of a black 96-well plate in order to image their ventral side. The fluorescence of the tadpoles is then measured by measuring the total fluorescence in each well. Following the fluorescence reading, tadpoles are then euthanised, frozen, and disposed of according to regulations at each respective laboratory.

2.4. Replication

One test is composed of three independent and valid runs using 20 tadpoles/group (see figure 2). Each run must be performed using independent solutions and spawn. The raw data for a given test chemical are obtained by pooling together the data from the three runs to obtain n=60 fluorescence values in each experimental group.

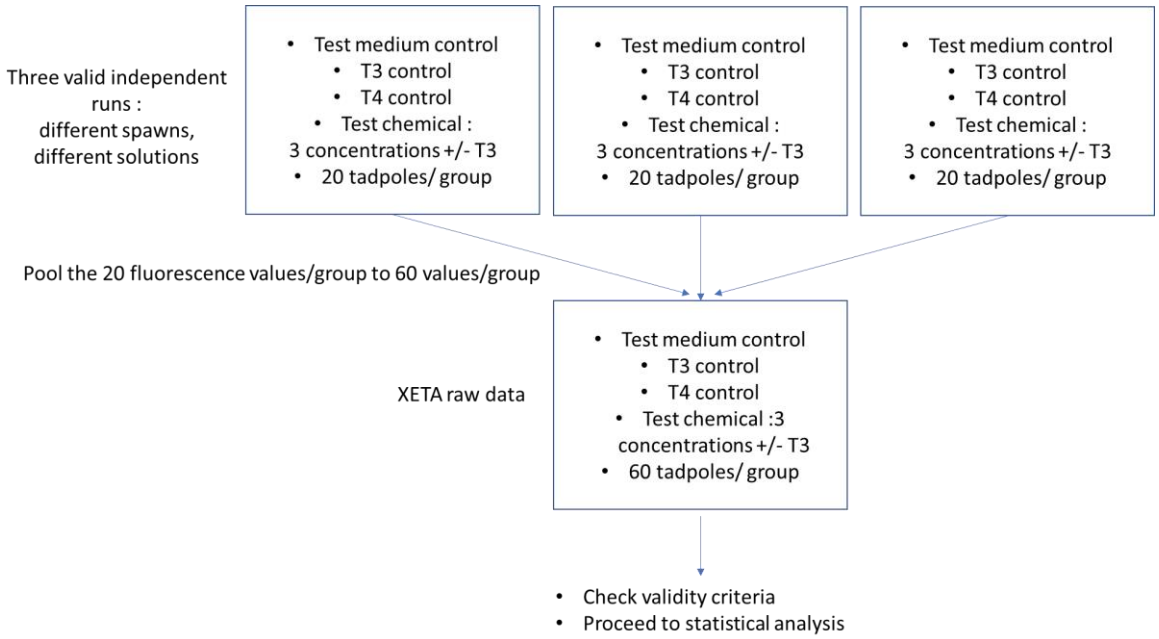

Figure 2: Overview of the XETA.
All chemical are tested with or T3 ("+/- T3"). The three concentrations test design shown here is the test design proposed in the draft test guideline. For the validation process a five concentrations test design has been used.

3. Validation phase I

3.1. Specific goals of phase I

1) Each laboratory performed a "calibration" experiment to establish whether performing laboratories obtained the expected dynamic range of fluorescence with seven concentrations of T3. The calibration was also intended to help laboratories adjust their method(s) and the settings of the spectrofluorimeter to obtain the optimal sensitivity in fluorescence readings.
2) A ring-test protocol was conducted within the three participating laboratories and intra- and inter-laboratory variability was determined. Four test compounds that display different mechanisms of action within the thyroid system (two agonists, one antagonist, and one negative) were tested to obtain sufficient information on the reliability, reproducibility, and sensitivity of the assay. T3 served as the positive control and all test chemicals were tested with and without T3.
3) Performance was compared between the participating laboratories based on the analysis of the results by a statistician. Reliability, reproducibility within and across laboratories, and sensitivity of the assay were determined.
4) The minimal number of tadpoles that can be utilized in the test through statistical evaluation of the phase I data was determined.

3.2. Overview of Test Conditions

The lead laboratory, WatchFrog provided participating laboratories with adult frogs, which were used to breed tadpoles for use in the experiments.

For the phase I validation, T4 and TRIAC (3,5,3'-Triiodothyroacetic Acid) served as thyroid receptor agonists, PTU (Propylthiouracil) an inhibitor of key enzymes in the thyroid hormone pathway, and cefuroxime as the inert (no variation of fluorescence was expected compared to untreated tadpoles).

All four chemicals were tested at five concentrations (indicated in the table below) and were tested with and without T3 (3.25 µg/L) added to the wells. The chemicals used by all three laboratories were from the same batch and lot number.

Preliminary experiments have been performed in the lead laboratory using the proposed substances to determine appropriate testing concentrations. For PTU and cefuroxime, both substances were

determined to be soluble and not presenting acute toxicity for the tadpoles while exposed to a concentration of 100 mg/L for 72h at 21°C. The highest concentration of the concntration range for the XETA was therefore set to 100 mg/L in accordance to the Amphibian Metamorphosis Assay guidance setting the highest concentration to 100 mg/L, the solubility limit or the Maximum Tolerated Concentrations (MTC). For T4 and TRIAC, using the Amphibian Metamorphosis Assay guidance for selecting concentrations would have led to test only concentrations inducing the maximal fluorescence observable, therefore we selected a range of concentration expected to induce a maximal effect for the highest and no effect for the lowest.

The testing began with a calibration experiment. The goal of the calibration step was to ensure that all laboratories obtained the same amplitude of fluorescence response in simple conditions and that no major differences were found when using different equipment. Because laboratories used a different type of fluorescent reader, it was crucial to demonstrate that both types of readers provided a suitable dynamic concntration response with T3 before testing of other chemicals began.

The calibration experiments determined:

1. The variability between controls
2. The amplitude and variability of T3 induction with a concentration of 3.25 µg/L
3. The amplitude of induction of fluorescence using seven increasing concentrations of T3

Once a laboratory demonstrated its ability to run the calibration experiments with an expected dynamic concentration response for T3, it then obtained the agreement of the lead laboratory to begin to test the chemicals of interest.

The following conditions for the test are summarized in the table below:

Test Animal		*Xenopus laevis* embryo
Exposure period		Stage NF45 for 3 days (72h)
Criteria for selecting test individuals		Primary criterion will be developmental stage and health of animal (alive and no malformations)
Calibration Experiment		T3 (3.25 mg/L; 0.65 mg/L; 32.5 µg/L; 6.5 µg/L; 3.25 µg/L; 0.65µg/L; 0.32µg/L)
Test medium Control		Test medium only (basal fluorescence)
T3 Control		T3 (3.25 µg/L)
T4 Control		T3 (3.25 µg /L) + T4 (10 mg/L)
Concentration of test substances	T4	10.0, 3.0, 1.0, 0.3, 0.1 mg/L
	TRIAC	1.0, 0.1, 0.01, 0.001, 0.0001 mg/L
	PTU	100, 30, 10, 3, 1 mg/L
	Cefuroxime	100, 30, 10, 3, 1 mg/L
Renewal		Every 24 hours for 72 hours (2 renewals)
Endpoints		Total fluorescence of individual tadpole
Samples per concentration		10 tadpoles per well (6-well plate) x 2 wells (total of 20 tadpoles per concentration for each run). 60 tadpoles per concentration in total
Volume of test medium		8 mL per well
Test medium		FETAX medium
Measurement time		72 hours post exposure
Runs (the number of experiments performed for each chemical)		Three runs (experiments will be run 3 times for each compound). Each run utilizes different spawn and test solutions independently prepared

Table 1: Conditions of the Xenopus Eleutheroembryonic Thyroid Signalling Assay

A sample assay design included the following chemicals and test concentrations as outlined in Table 2 below. *Note: Not more than two chemicals have been run per assay per week.*

Test Group	**Exposure Medium Contents**	**Number of wells (10 embryos/well)**
Groups without T3		
Test medium control	FETAX medium	2
Test chemical*	Test medium + test chemical (5 concentrations)	2 per concentration (10 per chemical)
Groups with T3		
T3 control	T3	2
T4 control	T3 + T4	2
Test chemical* with T3	Test chemical (5 concentrations) + T3	2 per concentration (10 per chemical)

* T4, TRIAC, PTU, or Cefuroxime

Table 2: Assay Design with one test chemical

3.2.1. Test Medium

FETAX medium was chosen as the test medium because this medium is used for the standardized ASTM Frog Teratogenicity Assay in Xenopus (ASTM 2012) and has been shown to be suitable for the growth of Xenopus embryos and tadpoles.

The FETAX medium has the following composition:

- 625 mg/L NaCl
- 96 mg/L $NaHCO_3$
- 30 mg/L KCl
- 15 mg/L $CaCl_2$
- 60 mg/L $CaSO_4$ $2H_2O$
- 75 mg/L $MgSO_4$
- pH is adjusted between 7.5-8.0 with a solution of NaOH 1N

3.2.2. Test validity

XETA experiments were judged valid during the validation exercise if:

- The T3 control group shows a statistically significant increase of the mean tadpole fluorescence of at least 20% compared to the test medium control group.
- The T4 control group shows a statistically significant increase of fluorescence compared to the T3 control group.
- The mortality does not exceed 20% in each experimental group (i.e. each group of 20 tadpoles corresponding to a test concentration or a control group)

3.2.3. Training

Two of the naïve laboratories that participated in the phase I validation were initially trained by staff from the WatchFrog laboratory before conducting the validation to ensure that each of the steps of the protocol were clearly understood.

3.2.4. Equipment

The following 96-well plate spectrofluorimeters were used by the participating laboratories:

Laboratory	**Japan**	**USA**	**France**
Spectrofluorimeter	Infinite 200 Pro (Tecan)	Synergy 2 (Biotek)	Infinite 200 Pro (Tecan)

Table 3: Spectrofluorimeters used for phase I

WatchFrog and the Japanese laboratories used the infinite F200 Pro (Tecan), a 96-well plate spectrofluorimeter using optical filters to generate the correct excitation and emission wavelengths. The US laboratory used a different reader, a Biotek Synergy 2 microplate reader. This reader works with a monochromator (an optical device that generates a narrow band of wavelengths of light) instead of optical filters.

3.3. Results of the phase I

3.3.1. Determination of statistical method

The lead laboratory proposed a statistical method in the XETA phase I SOP to allow the participating laboratories to analyse the data during the data generation phase. This was necessary to allow the laboratories to check that the control values satisfied the test validity criteria. After the completion of the experiment by the three laboratories, the leading laboratory collected all raw data and submitted them for evaluation by an independent expert statistician (John W. Green).

The primary objectives of the statistical analyses were to first determine an appropriate method for analysing data generated through XETA, and to then use the recommended method to analyse the data and generate results. The reliability, reproducibility, and sensitivity of the assay were determined through statistical analysis for the positive control (T3) and the four test compounds. Analyses also compared the fluorescence in unspiked (without T3 added) and spiked (with T3 added) modes for all test chemicals. Reliability (whether several laboratories could obtain similar results with the same protocol) and reproducibility (whether the same results could be obtained by the same laboratory when performed on different days) were also determined.

3.3.2. Power analysis

A first power analysis was done after the completion of phase I, table 4 contains a brief indication of the power of the Williams, Dunnett, Jonckheere-Terpstra, and Dunn tests under conditions of variability and experimental design similar to what was observed in the XETA data. There is little power advantage of 20 tadpoles per treatment group over 10. There is a 15-20% power increase in the 3-rep design compared to a 2-rep design. Under the Williams or Dunnett test, there is high probability of a significant test if there is a true 25% change or higher in fluorescence values compared to the control. There is inadequate power to detect a 10% change. The non-parametric Dunn test is seriously under-powered to detect even a 50% increase in fluorescence values. The Jonckheere-Terpstra test based on runs means has adequate power to detect a 50% or greater increase, though that was not apparent in the results of the ring test.

Concentration=6, reps=2, sample size=10					
REPS	PCT_EFF	Williams	JT	Dunnett	Dunn
2	10	40%	31%	24%	9%
2	15	69%	40%	49%	13%
2	25	97%	52%	90%	16%
2	50	100%	84%	100%	22%
Concentrations=6, reps=2, sample size=20					
REPS	PCT_EFF	Williams	JT	Dunnett	Dunn
2	10	41%	31%	24%	8%
2	15	69%	41%	48%	12%
2	25	98%	53%	93%	16%
2	50	100%	83%	100%	21%
Concentrations=6, reps=3, sample size=10					
REPS	PCT_EFF	Williams	JT	Dunnett	Dunn
3	10	60%	41%	39%	15%
3	15	90%	57%	75%	28%
3	25	100%	71%	99%	38%
3	50	100%	96%	100%	49%
Concentrations=6, reps=3, sample size=20					
REPS	PCT_EFF	Williams	JT	Dunnett	Dunn
3	10	61%	41%	42%	16%
3	15	90%	56%	77%	28%
3	25	100%	68%	99%	37%
3	50	100%	96%	100%	52%

Table 4: Selected Power of Statistical Tests

Concentrations: number of concentration tested (including the control), Reps : number of runs, sample size : number of tadpoles at each test concentration)

Analysis of the ring tests results and power simulations suggested that Williams' test is the preferred statistical test to use, possibly after a normalizing, variance stabilizing transformation, taking runs into account. This assumes the data are consistent with a monotone concentration-response. Where the data are seriously inconsistent with monotonicity, Dunnett's test can be used, again possibly after a normalizing, variance stabilizing transformation, taking runs into account. There are no known non-parametric methods that take the runs properly into account that also have adequate power to detect effects if present unless replication is substantially increased.

3.3.3. Trimming

A trimming was performed prior statistical analysis by omitting in each run the highest and lowest 10% of the fluorescence values for each control group and each concentration of T3-spiked and unspiked test concentration (i.e. omitting the two highest and two lowest values of each group of 20 values). This trim is intended to remove values that arise in the XETA from several events including missorted tadpoles (abnormal size or pigmentation), misplaced tadpoles in the 96-well plate (tadpoles upside down or on their side), tadpoles dying after anaesthesia (the death of the tadpole leads to the release of autofluorescent material from the gallbladder) and wells containing no tadpoles.

3.3.4. False positive rate

The statistical tests used control the false positive results at the 5% level.

3.3.5. Establishing a decision logic

The XETA is intended to be used as a screening assay. The XETA result will likely influence the conduct of any additional testing. In this context, the need of a decision logic to establish if the results of the assay may be interpreted as either positive, negative (or possibly equivocal) for a given chemical was discussed during the VMGeco in 2015. Several possible decision logics were proposed including taking into account the statistical significance and/or a threshold for percentage increase in fluorescence.

After having analysed the phase II data and re-analysed the phase I data with the revised statistical approach, the following decision logic was selected combining statistical significance and a fluorescence variation threshold. This threshold has been set to 12 %. The 12% threshold came from the power analysis which showed approximately 70 and 90% power to detect 10 and 15% effects, respectively. Interpolation indicates approximate 80% power to detect a 12% effect. An 80% power is widely accepted as adequate. (See, for example, OECD 2006 or Green *et al*. 2018.) In the draft test guideline a design with three test concentrations plus control is proposed, the power to detect a 12% effect was specifically calculated and found to exceed 80% for all concentration-response shapes simulated previously. The adequation of this threshold was confirmed by the results for the two validation

phases showing no groups of tadpoles treated with a reference inactive chemical giving a statistically significant variation over 12%.

A decision logic was developed for the XETA to provide logical assistance in the conduct and interpretation of the result of the bioassay (see flow chart in figure 3). This decision logic is based on three valid runs pooled for statistical analysis (see figure 2). A test chemical is considered to give a positive result in the XETA if at least one concentration tested including the highest is active in T3-spiked and/or unspiked mode.

-In unspiked mode an active concentration is defined as a concentration giving a statistically significant fluorescence increase of 12% or greater compared to the test medium control.

-In T3-spiked mode an active concentration is defined as a concentration giving a statistically significant fluorescence increase or decrease of 12% or greater compared to the T3 control.

Fluorescence decreases in unspiked mode are not expected as the tadpole does not synthetize its own thyroid hormone at this development stage. This could be easily illustrated by the fact that testing a specific antagonist of the thyroid receptor at high concentration does not lead to a decrease in fluorescence in unspiked mode (see NH3 (a TR specific antagonist) results during phase II). If a statistically significant fluorescence decrease >12% is observed in unspiked mode, it should be considered to repeated the XETA using a lower concentrations range or performing another test.

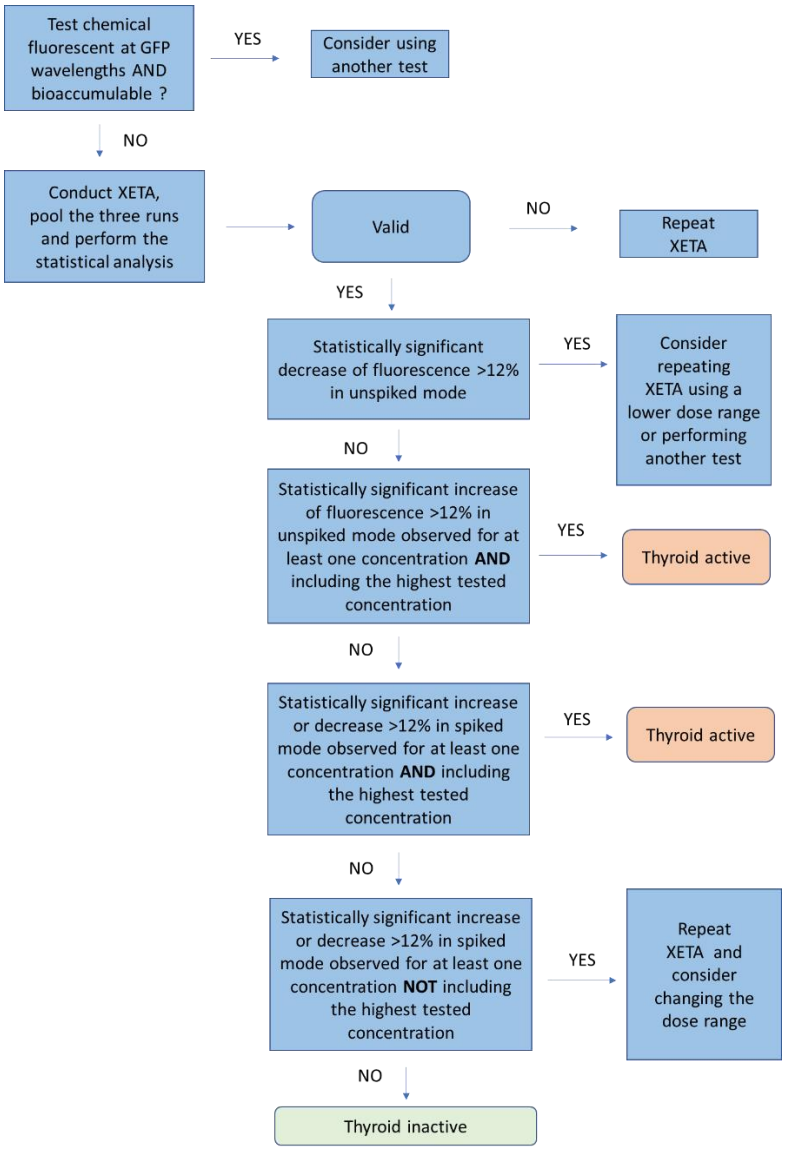

Figure 3: Decision logic for the conduct of the XETA

3.3.6. Establishing NOEC and LOEC

The result of the XETA is intended to be a classification of the test chemicals into potentially "thyroid active" or "thyroid inactive". The results of the XETA are expressed here in terms of LOEC and NOEC to allow the comparison of the results between the participating laboratories.

The LOEC would then be defined as the lowest concentration found to be active either in unspiked or spiked mode.

The NOEC would then be defined as the concentration tested immediately below the LOEC.

3.4. Results of Analyses

The results presented here are the results obtained with the statistical approach and decision logic determined after the analysis of the phase I data and phase II datasets.

3.4.1. Interlaboratory control comparison

The CVs for the same chemical were compared across laboratories. The mean used to calculate the CV was the control mean for the unspiked experiments and the T3 control mean for the spiked experiments. Two methods were used. For the dataset from a single lab, chemical and phase, after the trimming, an ANOVA was done with Rep and Rep*Treatment as random effects and Treatment as the sole fixed effect. The ANOVA removed the treatment effects. The pooled variance of the treatment means from this model were computed, as this is the variance used in statistical tests. Two versions of CV were computed, one using the standard error (square-root of the reported variance, CVSE) and the other using the standard deviation (square-root of reported variance multiplied by the degrees of freedom, CVSD). The former is more applicable to the statistical tests, while the latter is more traditional.

In the unspiked mode (Table 5), the mean CVSEs from the three laboratories are close, ranging from 10 to 12.5. The French laboratory was the most consistent, as the CVs ranged from 9.1 to 16.4 across the four chemicals tested. This compares to ranges of 4.2 to 16.9 for Japan, and 3.9 to 20 for France and USA. No chemical-related pattern was apparent in CVs.

Laboratory	Chemical	StdErr	DF	Ctrl mean	CVSE	CVSD	mean CVSE	mean CVSD
FRANCE	CEF	498.1	10	3533.4	14.1	44.6	12.2	38.6
	PTU	268.1	10	2959.5	9.1	28.6		
	T4	330.3	10	3517.9	9.4	29.7		
	TRIAC	509.8	10	3109.2	16.4	51.8		
JAPAN	CEF	154.6	10	3651	4.2	13.4	12.6	40,0
	PTU	649.4	10	3843	16.9	53.4		
	T4	448.6	10	3099.5	14.5	45.8		
	TRIAC	686.3	10	4582	15	47.4		
USA	CEF	1.5	10	37.6	3.9	12.3	10.0	31.6
	PTU	3.9	10	43	9	28.3		
	T4	2.9	10	39.6	7.3	23		
	TRIAC	8.5	10	42.4	20	63.1		

Table 5: Distribution of CV (unspiked mode). StdErr : standard error, DF : degree of freedom, Ctrl mean : unspiked control mean.

In spiked mode, table 6, the mean CVSEs from the three laboratories are close, ranging from 7.6 to 10. The French laboratory was the most consistent, as the CVs ranged from 6.3 to 11.5 across the four chemicals tested. This compares to ranges of 3.2 to 15.5 for Japan, and 3.6 to 16.2 for USA. No chemical-related pattern was apparent in CVs.

Laboratory	Chemical	StdErr	DF	Ctrlmean	CVSE	CVSD	mean CVSE	mean CVSD
FRANCE	CEFT3	518.5	10	4651	11.1	35.3	8.8	27.9
	PTUT3	286.7	10	4567.4	6.3	19.9		
	T4T3	314.7	10	4925.6	6.4	20.2		
	TRIACT3	576.3	10	5013.4	11.5	36.4		
JAPAN	CEFT3	151.5	10	4794	3.2	10	10,0	31.6
	PTUT3	849.3	10	5471.4	15.5	49.1		
	T4T3	313.9	10	4232.1	7.4	23.5		

	TRIACT 3	827.3	10	5957.9	13.9	43.9		
USA	CEFT3	2.1	10	49.3	4.2	13.2	7.6	24.3
	PTUT3	4.2	10	62	6.7	21.3		
	T4T3	1.8	10	51.3	3.6	11.4		
	TRIACT 3	9.7	10	59.7	16.2	51.3		

Table 6: Distribution of CV (spiked mode).

StdErr : standard error, DF : degree of freedom, Ctrl mean : T3 control mean.

3.4.1. Interlaboratory test chemical CV comparison

The variability in each experimental group for each chemical was assessed by calculating the coefficient of variation (100*SD/Mean) associated with the mean of the fluorescence values after the 10% trim (i.e. 48 fluorescence values in each group). The table 7 and figure 4 give an overview of the results. The CV for individual experimental group ranged from 8 to 28%, there is no laboratory effect, the mean for all CV are nearly the same in spite of the differences in apparatus used to read the experiments (18% for France, 17% for Japan and USA).

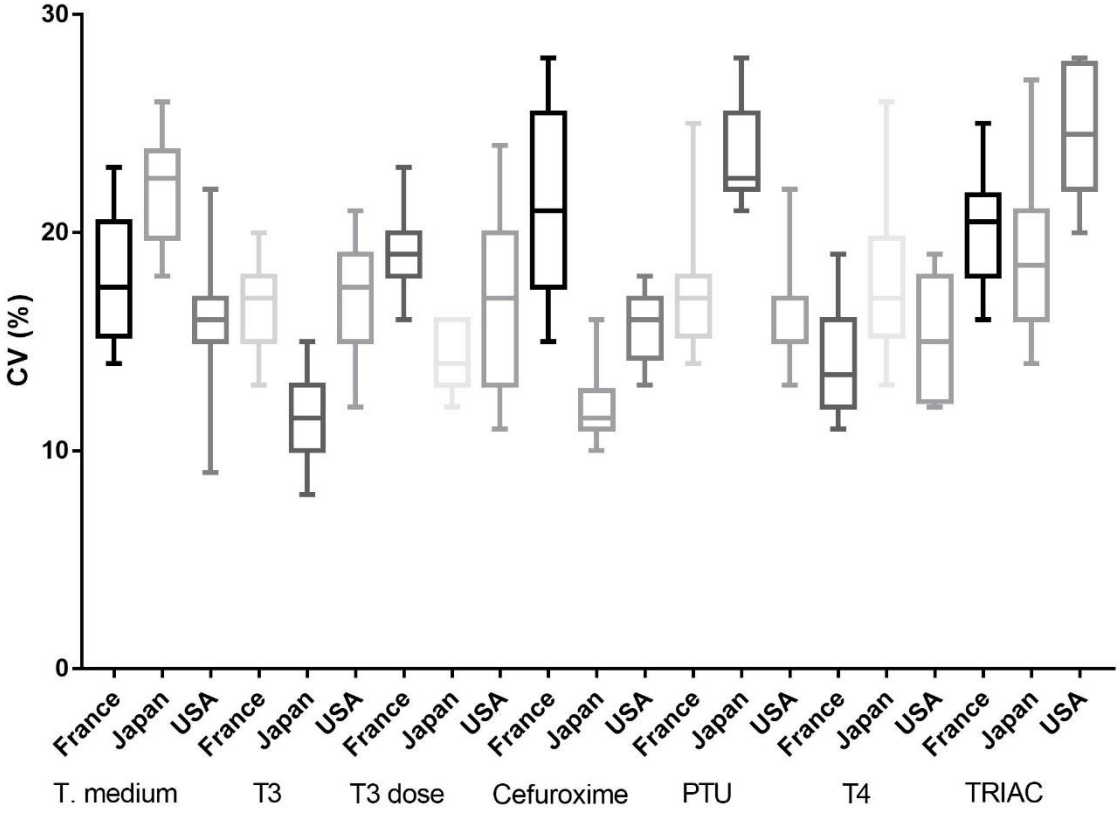

Figure 4: Overview of the CV in each experimental group of the phase I.
"T3" refers to the calibration experiments including 11 repeats of the T3 control group. "T3 dose" refers to the T3 concentration response experiments.

		T3	Test medium										Mean of test medium CV
France	Mean	4879	2930	3363	2888	2986	2931	3180	2897	3004	3149	3038	3006	
	CV	15%	22%	17%	16%	15%	19%	21%	18%	19%	16%	14%	23%	18%
Japan	Mean	5822	3721	4041	3732	3822	3729	4075	3519	3754	3693	3964	3632	
	CV	26%	22%	23%	24%	25%	23%	22%	19%	22%	23%	19%	18%	22%
USA	Mean	72	52	55	49	53	53	50	48	54	52	48	49	
	CV	15%	9%	17%	15%	16%	17%	21%	17%	22%	16%	15%	15%	16%

		Test medium	T3										Mean of T3 control CV
France	Mean	3394	5791	6347	5975	5869	6063	6159	5946	5988	6213	6243	5822	
	CV	15%	18%	18%	17%	17%	18%	19%	17%	13%	15%	20%	15%	17%
Japan	Mean	3720	5276	5402	4798	4863	5180	5196	4567	4823	5063	5142	4728	
	CV	15%	11%	15%	13%	10%	11%	13%	8%	13%	10%	12%	10%	12%
USA	Mean	48	65	73	64	65	70	63	64	69	64	68	68	
	CV	18%	12%	17%	15%	13%	20%	19%	15%	18%	19%	17%	21%	17%

		Test medium	T3	T3 0.32µg/L	T3 0.65µg/L	T3 3.25µg/L	T3 6.5µg/L	T3 32.5µg/L	T3 0.65mg/L	T3 3.25mg/L	Mean of CV
France	Mean	3559	5716	3490	3512	5731	6297	6889	6960	7751	
	CV	18%	21%	23%	18%	20%	16%	19%	18%	20%	19%
Japan	Mean	3496	4580	3502	3794	4911	5154	5816	5960	6544	
	CV	16%	13%	15%	13%	12%	15%	16%	16%	13%	14%
USA	Mean	45	63	46	47	62	72	87	86	87	
	CV	20%	17%	14%	11%	24%	11%	13%	24%	20%	17%

			-T3						+T3						Mean of CV
			Test medium	C1	C2	C3	C4	C5	T3	C1	C2	C3	C4	C5	
CEFUROXIME	France	Mean	3533	4142	4062	3351	3360	3469	4651	5399	4899	4902	5178	5118	
		CV	24%	22%	19%	28%	23%	26%	28%	15%	17%	17%	20%	19%	**22%**
	Japan	Mean	3651	3811	3599	3612	3428	4056	4794	4386	4971	4733	4775	4455	
		CV	12%	16%	11%	13%	12%	12%	13%	10%	10%	11%	11%	11%	**12%**
	USA	Mean	38	41	37	37	38	41	51	48	51	50	49	48	
		CV	17%	14%	15%	16%	18%	17%	13%	14%	17%	15%	18%	16%	**16%**
PTU	France	Mean	2960	2894	2924	2899	2843	3447	4567	4461	4963	4697	5132	5520	
		CV	21%	18%	25%	18%	16%	18%	18%	15%	16%	16%	14%	14%	**17%**
	Japan	Mean	3843	4256	3959	3801	3676	4930	5471	5181	5681	5972	5936	6571	
		CV	26%	28%	23%	22%	22%	24%	21%	22%	23%	22%	21%	26%	**23%**
	USA	Mean	44	47	43	43	45	49	61	66	70	69	69	75	
		CV	13%	17%	15%	15%	15%	15%	17%	16%	22%	22%	15%	13%	**16%**
T4	France	Mean	3518	3911	5489	7449	7372	8104	4926	5065	6479	7407	7571	7476	
		CV	19%	16%	17%	16%	14%	11%	13%	12%	12%	15%	13%	12%	**14%**
	Japan	Mean	3099	4142	5246	6460	6134	6654	4232	5128	6093	6097	6447	6232	
		CV	19%	20%	16%	17%	17%	20%	15%	15%	16%	26%	13%	18%	**18%**
	USA	Mean	40	45	54	65	66	66	51	52	61	65	66	63	
		CV	18%	15%	17%	13%	12%	18%	18%	19%	12%	13%	15%	12%	**15%**
TRIAC	France	Mean	3109	3321	3322	5410	6089	6887	5013	4437	4809	5749	6400	5814	
		CV	25%	20%	23%	22%	18%	16%	18%	21%	21%	17%	21%	19%	**20%**
	Japan	Mean	4582	4921	5150	6583	7326	8833	5958	5784	6669	7312	7589	7454	
		CV	21%	21%	21%	19%	14%	16%	18%	27%	22%	18%	16%	15%	**19%**
	USA	Mean	42	49	55	71	75	78	60	64	71	74	79	78	
		CV	28%	27%	24%	21%	28%	24%	28%	22%	20%	26%	22%	25%	**25%**

Table 7: Overview of the mean and CV in each experimental group of the phase I.

3.4.2. Control groups

One of the objectives of the statistical analysis was to determine whether the experimental design and statistical tests used have the power to detect a significant difference between test medium control and T3 control, and a difference between T3 control and T3+T4 control. All of these differences were found statistically significant in every experiment.

The T3+T4 treatment in the T4 control group induced a high level of fluorescence close to the plateau of the concentration response curve. The fluorescence quantification of these tadpoles never reached the saturation level of the spectrofluorimeters with the applied settings showing that every laboratory could effectively quantify the highest levels of fluorescence reached by the tadpoles.

Criteria of validity for the XETA included reaching a 20% increase in fluorescence in the T3 control group. All experiments included in the dataset satisfy this criterion with inductions ranging from 24 to 62% for USA, 23 to 75% for France and 21 to 63% for Japan (Figure 5).

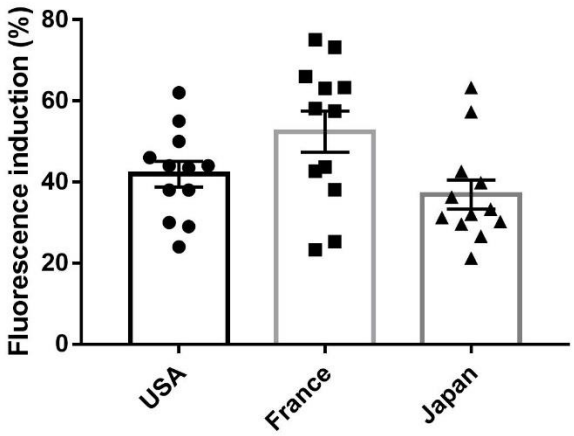

Figure 5 : Fluorescence induction observed in the T3 control groups over the 12 phase I experiments. The mean, SEM and the values for every run are presented.

3.4.3. Calibration

Test medium control (FETAX)

Two runs (Japan and USA) or three runs (France) of an experiment including 11 groups of tadpoles treated only by the test medium were performed. The data from the Japanese, USA, and France laboratories is included in figure 6 below showing the mean and SEM of the fluorescence for each test medium group in each laboratory. A T3 control group was included in each experiment to ensure that an expected difference in fluorescence could be observed between each test medium group and the T3 control. The analysis shows all test medium group to be statistically different from the T3 control.

FETAX served as a blank for the assay and all three laboratories obtained similar background fluorescent readings. Variances in actual fluorescent readings were due to differences in the equipment used to measure fluorescence. Other factors are also likely to have contributed to differences across laboratories, but the relative increases in fluorescence of the positive controls and test chemicals were similar across laboratories as were the coefficients of variance. Therefore, the calibration experiment demonstrated that the variability between controls was acceptable.

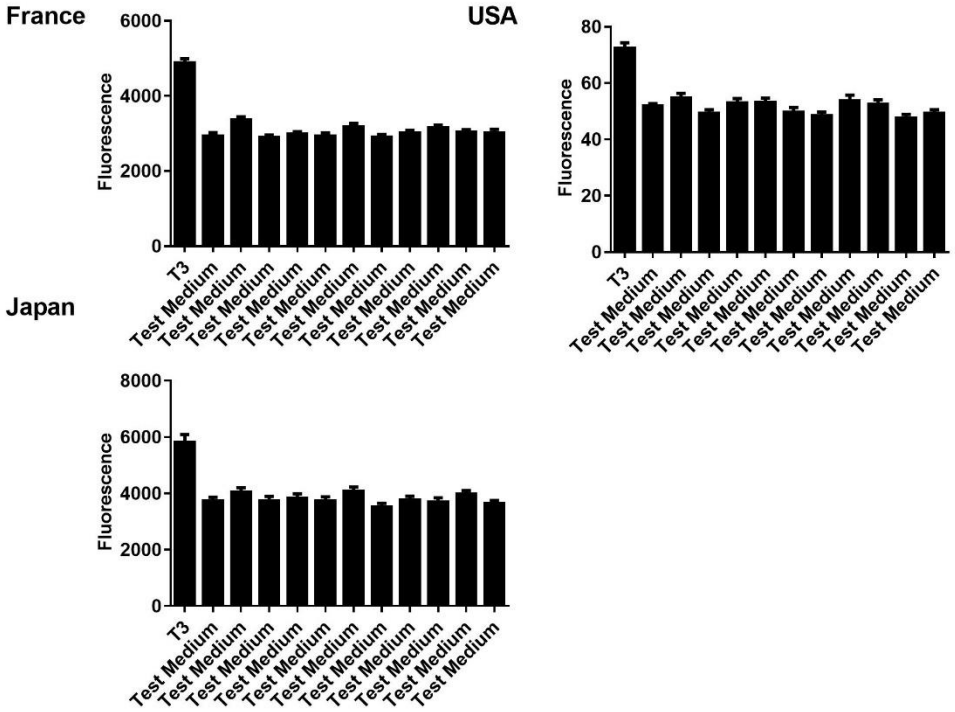

Figure 6: Mean and SEM of 11 test medium control groups.
The fluorescence values are expressed in Relative Fluorescence Units.

T3 (3.25 µg/L)

Two runs (Japan and USA) or three runs (France) of an experiment including 11 groups of tadpoles treated by 3.25 µg/L of T3 were performed. The data from the Japanese, American, and French laboratories are included in figure 7 below showing the mean and SEM of the fluorescence for each T3 group in each laboratory. A test medium control group was included in each experiment to ensure that an expected difference in induction could be seen between each T3 group and the test medium control. The analysis shows all mean T3 response to be statistically different from the test medium control.

The T3 (3.25 µg/L) calibration data allowed for a statistical determination of the minimum number of tadpoles needed in the experiment to detect a significant change in fluorescence. Analysis of this dataset indicated that decreasing the number of tadpoles in the experiment from 20 per concentration to 10 per concentration did not affect the results or sensitivity of the test. However, it was recommended that the protocol remain the same, using 20 tadpoles because the sensitivity of the experiment remains high when some of the tadpoles are lost due to experimental error or acceptable levels of

mortality during the exposure. The analysis also examined the need for three runs versus two runs in the protocol. This analysis determined that using only two runs decreased the sensitivity of the test by 15-20%. It was therefore recommended that the protocol continues to stipulate the use of at least three runs of each chemical tested in a given laboratory.

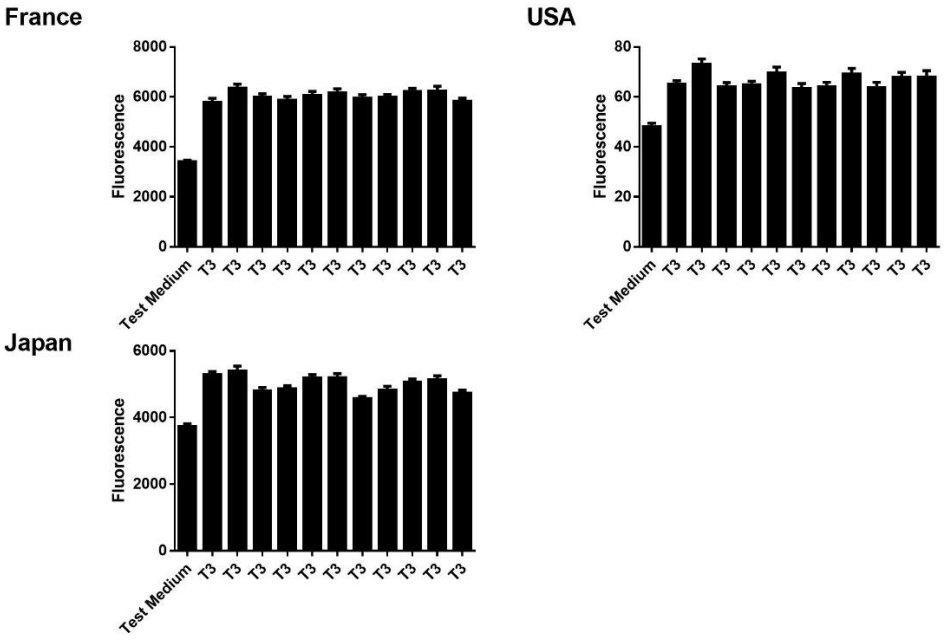

Figure 7: Mean and SEM of 11 T3 (3.25 µg/l) control groups.
The fluorescence values are expressed in Relative Fluorescence Units.

T3 Concentration-response

Three runs of seven concentrations of T3 were performed by the participating laboratories. The T3 data from the Japanese, American, and French laboratories are included in figure 8 below showing the mean and SEM of the fluorescence for each concentration of T3 in each laboratory.

The increase in fluorescence in each laboratory is clearly obtained with increasing concentrations of T3. All three laboratories obtained large and consistent increases (>40%) in fluorescence with every concentration of T3 above 0.65 µg/L. The statistical analysis shows all increases in fluorescence above 0.65 µg/L to be significant and over 12% (Table 8).

This result shows that the calibration experiments worked as intended. The three laboratories were all able to obtain a similar sensitivity in response to a range of T3 concentrations and in a consistent manner.

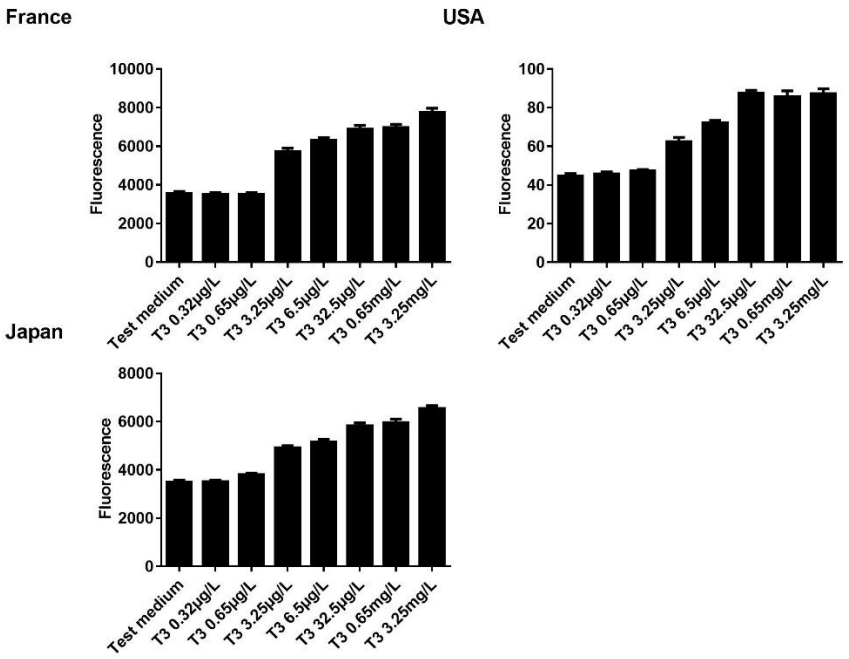

Figure 8: Mean and SEM of fluorescence intensities in the T3 concentration response calibration experiment. The fluorescence values are expressed in Relative Fluorescence Units.

Country	T3 (µg/l)						
	0.32	0.65	3.25	6.5	32.5	65	325
USA	-2	-1	61	77	94	96	118
France	3	6	40	62	96	92	96
Japan	0	9	40	47	66	70	87

Table 8: Percentage of fluorescence variations in the T3 experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

3.5. Results for substances

3.5.1. CEFUROXIME Results

Cefuroxime (CEF) was expected to be thyroid inactive. Figure 9 below shows the mean and SEM for each concentration of cefuroxime in each laboratory. Visual observation of the means plotted in this figure show a

rather flat response of the XETA assay to CEF, with occasional and random increases in fluorescence.

The statistical analysis found a significant 11% increase in unspiked mode at 100 mg/L in the results from the Japanese laboratory (Table 9). Please note that despite a power of 80%, William's test is not finding the 17%, 15% and 16% fluorescence variations in the French data to be significant.

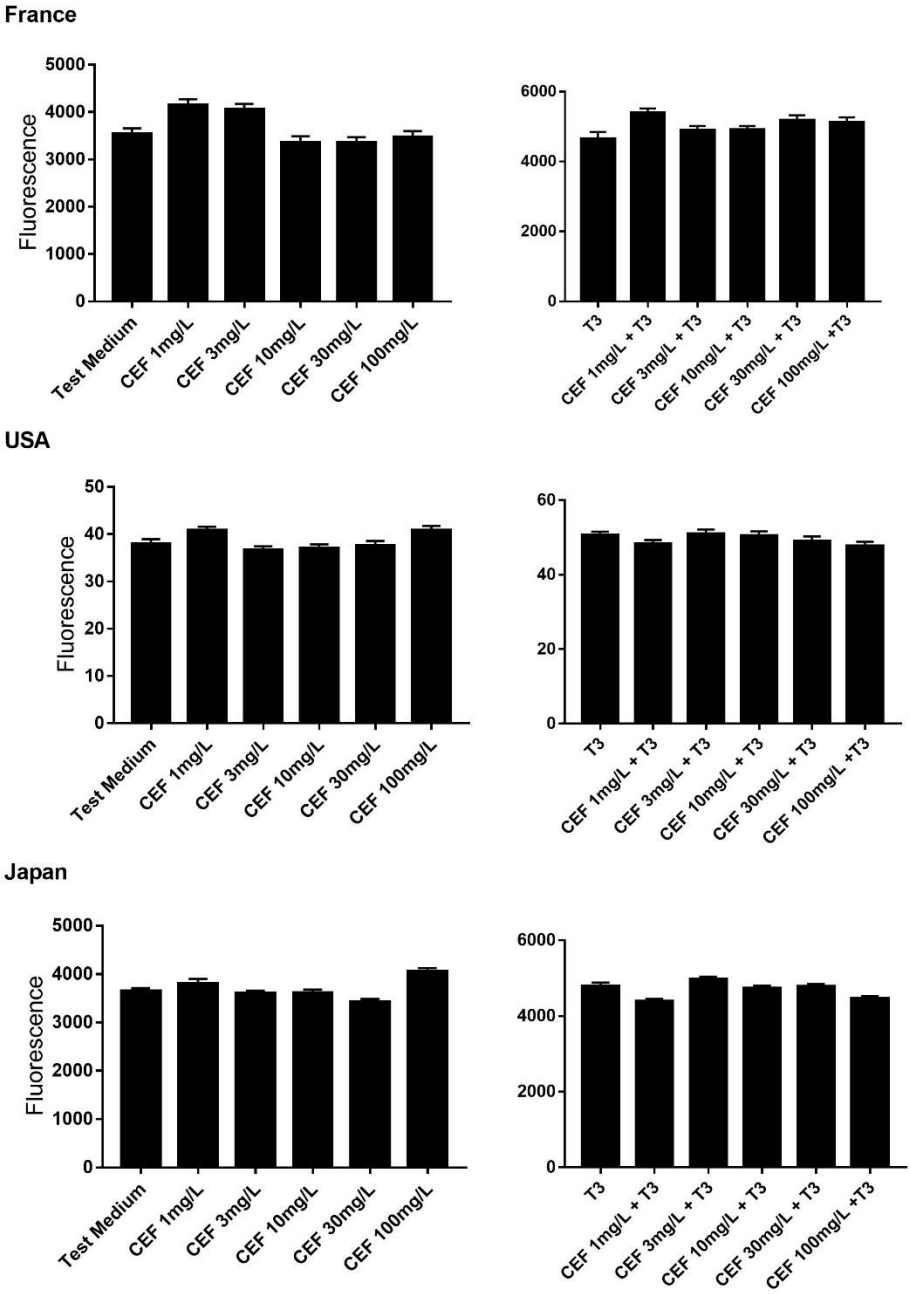

Figure 9: Mean and SEM of fluorescence intensities in the Cefuroxime experiment

Country	CEFU (mg/l)									
	-T3					+T3				
	1	3	10	30	100	1	3	10	30	100
France	17	15	-5	-5	-2	16	5	5	11	10
USA	7	-4	-3	-1	7	-4	1	0	-3	-6
Japan	4	-1	-1	-6	11	-9	4	-1	0	-7

Table 9: Percentage of fluorescence variations in the CEF experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

3.5.2. PTU Results

Figure 10 below shows the mean and SEM for each concentration of PTU in each laboratory. A consistent increase in fluorescence in each laboratory is clearly obtained with increasing concentrations of PTU in the presence of T3 and at the maximum concentration in the absence of T3.

The statistical analysis of the data shows a significant increase over 12% at 100 mg/L for France, USA and Japan in unspiked mode.

In spiked mode, a significant increase is detected for concentrations above 10 mg/L for France, 3 mg/L for Japan and 1 mg/L for USA. In the Japanese data only the 100 mg/L concentration induces a fluorescence increase over 12%, in the two other laboratories all statistically significant increases in fluorescence are equal or greater to 12% (Table 10).

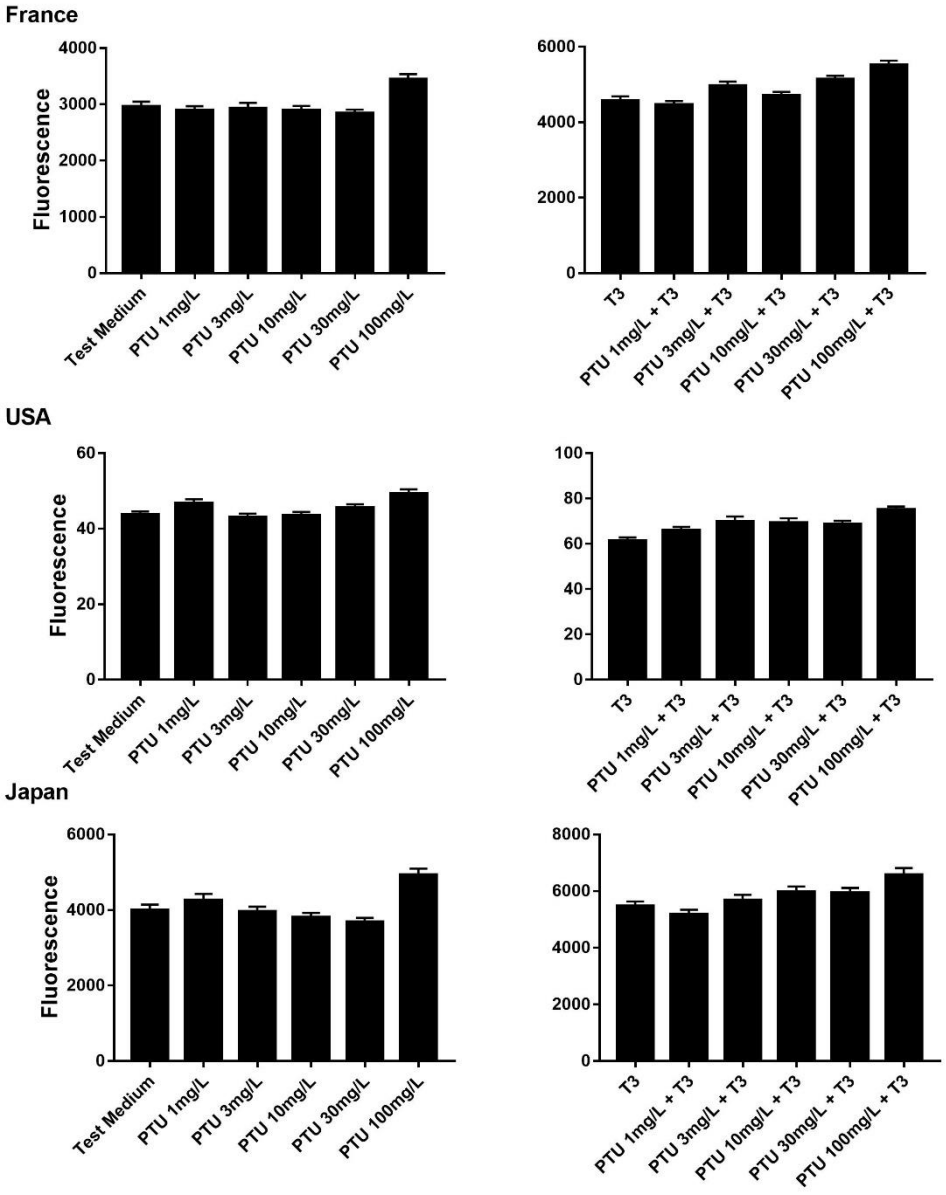

Figure 10: Mean of fluorescence intensities in the PTU experiment

Country	PTU (mg/l)									
	-T3					+T3				
	1	3	10	30	100	1	3	10	30	100
France	-2	-1	-2	-4	16	-2	9	3	12	21
USA	7	-2	-1	4	13	8	14	13	12	23
Japan	7	-1	-5	-8	23	-5	4	9	8	20

Table 10: Percentage of fluorescence variations in the PTU experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

3.5.3. T4 Results

There was strong consistency in results for T4. Figure 11 below shows the mean and SEM for each concentration of T4 in each laboratory. The increase in fluorescence in each laboratory is clearly obtained with increasing concentrations of T4.

For unspiked mode, France and USA found a significant increase in fluorescence over 12% for all concentrations above 0.01 mg/l. In the US laboratory a 13% increase at 0.01mg/l was observed at the lowest concentration and not found statistically significant. All concentrations were found significant and inducing over 12% of fluorescence increase in the Japan laboratory.

For spiked mode, all concentrations above 0.01mg/l were found to induce a significant increase over 12% in fluorescence by the three laboratories. In addition, the lowest concentration gave a significant increase over 12% in fluorescence in Japanese data (Table 11).

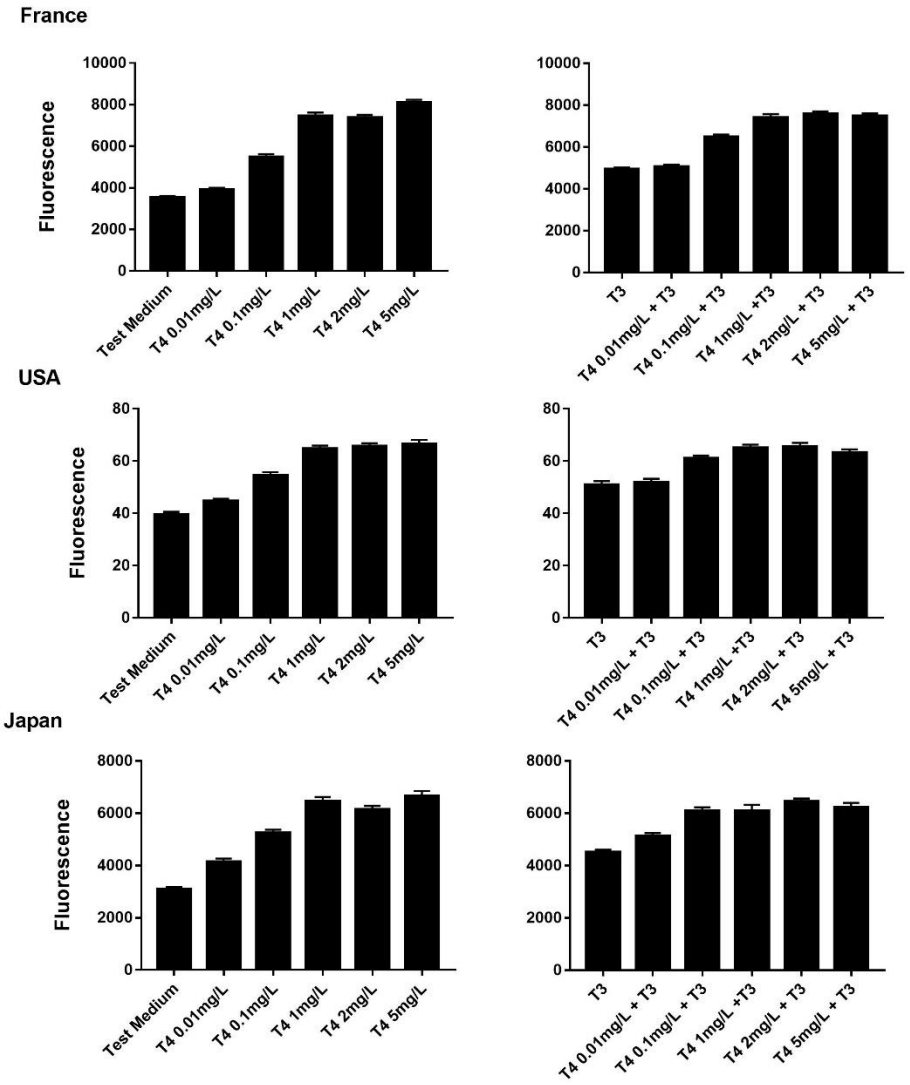

Figure 11: Mean of fluorescence intensities in the T4 experiment

Country	T4 (mg/l)									
	-T3					+T3				
	0.01	0.1	1	2	5	0.01	0.1	1	2	5
France	11	56	112	110	130	3	32	50	54	52
USA	13	38	63	66	68	2	20	28	29	24
Japan	34	69	108	98	115	14	35	35	43	38

Table 11: Percentage of fluorescence variations in the T4 experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

3.5.4. TRIAC Results

Figure 12 below shows the mean and SEM for each concentration of TRIAC in each laboratory. An increase in fluorescence in each laboratory is clearly obtained with increasing concentrations of TRIAC.

There was a general consistency in results across the three laboratories, but some differences were noted. In the US laboratory there was a significant increase in fluorescence over 12% above 0.0001 mg/L in unspiked mode, but only at 0.01 mg/L and above for the two other laboratories. A 16% increase at 0.0001 mg/l in the USA lab was observed and found not statistically significant.

For spiked test concentrations, significant increases equal or greater than 12% were found at 0.001 mg/L and above in the USA and Japan and at 0.01 mg/L and above for France (Table 12).

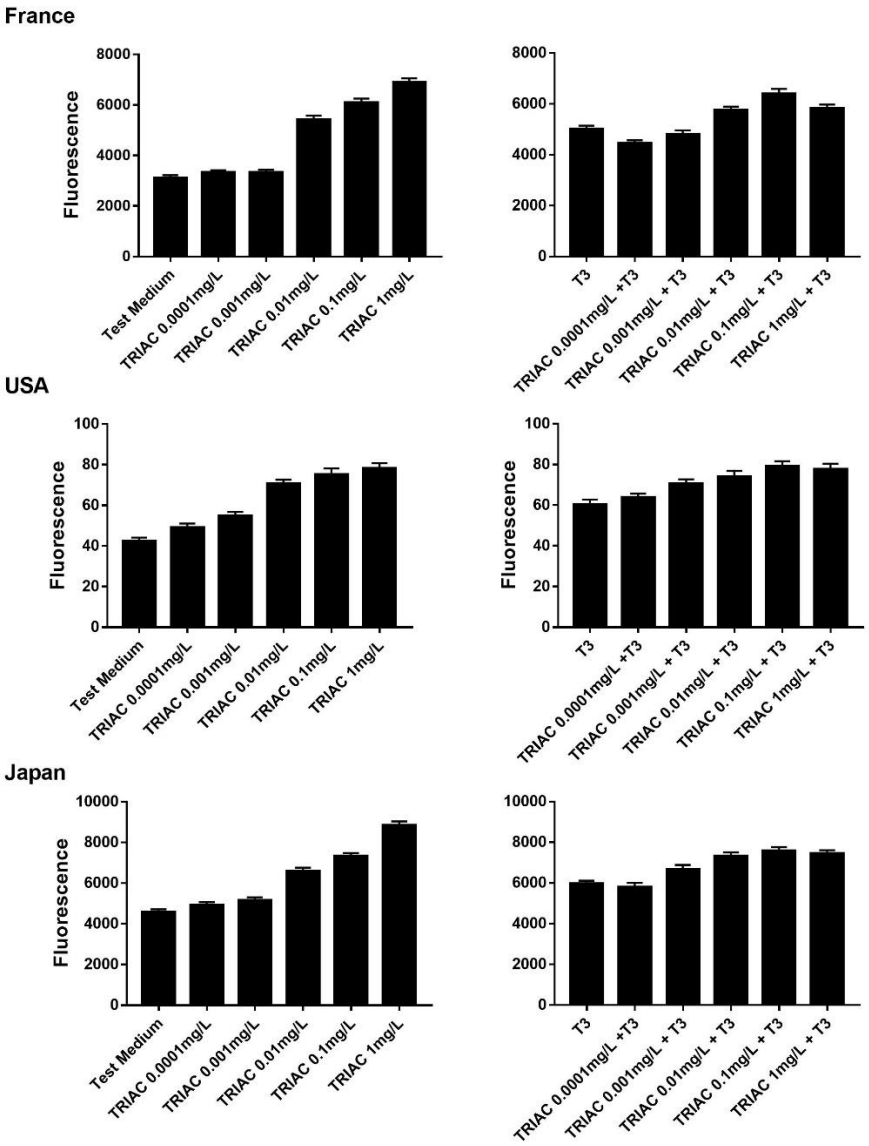

Figure 12: Mean and SEM of fluorescence intensities in the TRIAC experiment

Country	TRIAC (mg/l)									
	-T3					+T3				
	0.0001	0.001	0.01	0.1	1	0.0001	0.001	0.01	0.1	1
France	7	7	74	96	122	-11	-4	15	28	16
USA	16	30	66	77	84	6	17	23	31	29
Japan	7	12	44	60	93	-3	12	23	27	25

Table 12: Percentage of fluorescence variations in the TRIAC experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

3.6. Discussion

The analysed data indicate that the phase I validation effort using the XETA protocol produced reproducible and reliable data in the three laboratories across the controls and test chemicals. Despite a difference in the fluorescence reader used by one of the laboratories, all three laboratories were able to perform the assay with the expected sensitivity. The test chemicals and control substances also gave similar results across laboratories. Individual chemical results are discussed in detail below.

3.6.1. Calibrations

All three laboratories obtained consistent results and sensitivity in the calibration experiments, indicating that the laboratories were all able to run the T3 positive control and obtain similar sensitivities in response. The calibration experiments demonstrated that the variability between controls was acceptable for every laboratory. The calibration data also determined that three runs are needed. The calibration experiment shows a NOEC of 0.65 µg/l and a LOEC of 3.25 µg/l for T3 for every laboratory.

Based on these data, the phase II validation SOP included a method to evaluate the performance of the calibration experiment before entering the test chemical phase. The T3 concentration response range was modified to add concentrations between 0.65 µg/L and 3.25 µg/L in order to more precisely define the sensitivity of the XETA for T3.

3.6.2. TRIAC

There was a general consistency in results for TRIAC across the three laboratories. This substance clearly activates the fluorescent response as expected as TRIAC is a known synthetic TR ligand. The three laboratories found TRIAC to be a thyroid active substance with a LOEC of 0.01 mg/L for France and 0.001 mg/L for Japan and the USA.

3.6.3. PTU

In this assay PTU gives a clear increase in fluorescence across all laboratories at the higher concentrations. A response could be clearly observed in the spiked groups above 10 mg/L and in the unspiked groups at the maximum concentration (100 mg/L). The three laboratories identified PTU as a thyroid active molecule with LOEC of 30 mg/L for France, 10 mg/L for Japan and 3 mg/L for the USA.

PTU is known to interfere with thyroid signalling at different levels and leads to an increased fluorescence in this assay. PTU affects thyroid hormone production by several known actions. PTU inhibits thyroid peroxidase, preventing the formation of T4 in the thyroid gland. PTU is also known to affect thyroid hormone metabolism by acting as an inhibitor of the enzyme 5'-deiodinase (tetraiodothyronine 5' deiodinase or D1). D1 is involved in the deiodination of iodothyronines. It catalyses inner ring deiodination (IRD) which deactivates T3 and T4, and additionally, outer-ring deiodination (ORD) which converts T4 to the more active hormone T3. This action of PTU on D1 has been described in mammals but Xenopus D1 ORD activity has been shown to be essentially insensitive to inhibition by PTU (IC50> 1mM (152 mg/L)) (Kuiper et al. 2006). To date, no study on the effect of PTU on Xenopus D1 IRD activity are available.

At stage NF45 of Xenopus development, the thyroid gland is still developing and tadpoles do not yet produce their own TH, only maternal TH are present in the larvae. Inhibition of TPO is therefore expected to be without consequences on the fluorescence level.

There is a consistently observed increase in fluorescence in the XETA results suggesting that the net effect of PTU could be an overall increase in the amount of T3.

Moreover, the addition of T3 in the spiked groups leads to an increased sensitivity for PTU compared to the unspiked groups (the LOEC is 10 times lower in the presence of T3). This could suggest that the increase of fluorescence is the consequence of an inhibition of D1 IRD activity, with PTU inhibiting the inactivation of T3. Alternatively, if the Xenopus D1 IRD activity is insensitive to PTU, this could suggest an undescribed action of PTU in Xenopus on another enzyme leading to a decrease in T3 inactivation or clearance.

3.6.4. T4

The T4 results across all three laboratories were very consistent. T4 gave the expected results in this assay. As T4 is an endogenous thyroid hormone, and the precursor of the most biologically active form, T3, significant increases in fluorescence were expected at low concentrations. The three laboratories identified T4 as a thyroid active molecule with LOEC of 0.01 mg/l for France and Japan and 0.1 mg/L for USA.

This compound would be easily detected as a compound with clear thyroid effects in a screen using the XETA.

3.6.5. Cefuroxime (inert compound)

Cefuroxime (CEF) was selected as the inert compound for phase I validation of the XETA. To date, this antibiotic has no known effects on the thyroid system. No active concentrations were detected by the three laboratories, therefore CEF is categorized as a thyroid inactive molecule by the XETA.

3.6.6. Chemical analysis

The samples were frozen and shipped a chemical analysis company to perform Liquid Chromatography–Mass Spectrometry (LC-MS) on the test chemicals.

A major issue concerning the quantification of the compounds by LC-MS was reported by the chemical analysis company. After several unsuccessful attempts it appears that the high salinity of the FETAX medium renders the LC-MS quantification impossible. A clean up of the samples using reverse phase columns should have been performed prior to injection into the LC-MS apparatus. However, development and validation of this approach would have been required for each compound before the completion of the experiments.

After several attempts, the lead laboratory decided that no additional attempt would be made for the chemical analysis for the following reasons:

- No additional budget could be allowed for the development and validation of this specific method for the three substances.
- More importantly, most of the samples had been stored at -20°C for more than one year and a half and had been thawed at least one time, posing questions on the reliability of any results that could be obtained.

3.7. Conclusions

The XETA phase I result demonstrate that the assay provides the expected results with the chemicals tested and is reproducible across laboratories. Overall, the data generated in the three laboratories gave expected response profiles and the test chemical were correctly

classified into thyroid active or inactive in each laboratory. The issue with the chemical analysis during the phase I highlighted the need to adjust the test medium to facilitate chemical analysis.

4. Validation phase II

4.1. Protocols

4.1.1. Changes to protocol following phase I

The complete standard operating procedures used by the participating laboratories for the phase II could be found in Annex 2.

Water with appropriate characteristics for rearing tadpoles was used as the test medium instead of FETAX medium to avoid any interference with the chemical analysis.

In the phase I protocol, when a tadpole died during the exposure phase of the experiment it was removed from the experiment, resulting in an additional empty well in the 96-well plate used for fluorescence quantification. The statistical analysis performed after the phase I showed that the method recommended by the statistician for outlier removal does not remove all of these values in cases where several tadpoles died in the same group. To solve this issue, empty well positions corresponding to dead tadpoles were recorded during phase II experiments and corresponding empty well values were removed from the data prior to statistical analysis.

4.1.2. Training

For practical reasons, the phase II validation included a training video for new laboratories to understand how to perform critical steps in the protocol. The phase I validation included person-to-person training, but phase II results show that a training video is sufficient to convey the important points effectively to the participating laboratories.

4.1.3. General Experimental Design

The general experimental design (control groups, number of individuals, number of concentrations tested…) of phase II was identical to the phase I design.

The highest and lowest dilution of each chemical were analysed through analytical methods to ensure that exposure concentrations are correct within each laboratory.

4.1.4. Test medium

As for the Amphibian Metamorphosis Assay (OECD TG 231) the test medium is defined as "water with appropriate characteristic for rearing of tadpoles". Practically, it could be glass bottled mineral water or charcoal-filtered tap water permitting normal growth and development of *X. laevis* tadpoles. Because local water quality can differ substantially from one area to another, participating laboratories were advised that analysis of water quality should be undertaken to screen for potential contaminants and chemicals which could interfere with the assay. Special attention should be given that the water is free of copper, chlorine and chloramine; all of which are toxic to tadpoles, and free of known thyroid active molecules.

The participating laboratories chose the test medium detailed in the following table (table 13). As the Japanese and Portuguese laboratories had previously successfully used charcoal-filtered tap water to maintain and raise *X. laevis* tadpoles they chose not to perform any analysis of water. The French laboratory has previous extensive experience using Evian mineral water, therefore it was chosen as the test medium for the French phase II experiments. As Evian water is locally available in Belgium, the Belgium laboratory also chose this test medium.

Laboratory	Japan	Belgium	Portugal	France
Test medium	Charcoal-filtered tap water	Glass Bottled Evian mineral water	Charcoal-filtered tap water	Glass Bottled Evian mineral water

Table 13 : Phase II tests media

4.1.5. Provenance of tadpoles

Provenance of tadpoles for the participating laboratories is detailed in table 14. Adult Xenopus founders were sent by the French to the Japanese and Portuguese laboratories several months before the beginning of phase II experiments. These two laboratories planned to produce their own tadpoles for the validation phase II but a bacterial infection occurred in the Portuguese Xenopus colony and tadpoles had to be sent from France for the majority of the experiments. The French laboratory also shipped tadpoles every week to the Belgium team as this laboratory has no Xenopus facility.

Laboratory	Japan	Belgium	Portugal	France
Provenance of tadpoles	On-site reproduction	Tadpoles sent from France	Tadpoles sent from France and on-site reproduction	On site reproduction

Table 14 : Phase II provenance of tadpoles

4.1.6. Equipment

The following spectrofluorimeters were used by the participating laboratories:

Laboratory	Japan	Belgium	Portugal	France
Spectrofluorimeter	TriStar LB941 (Berthold)	Infinite 200 Pro (Tecan)	Synergy 2 (Biotek)	Infinite 200 Pro (Tecan)

Table 15 : Phase II spectrofluorimeters

All spectrofluorimeters used for phase II use fluorescent filters to obtain the correct wavelengths for GFP excitation and emission. The Infinite 200 Pro and the Synergy 2 models have been previously used during phase I.

4.1.7. Chemicals

Selection of chemicals

During the annual meeting in 2014 the VMGeco recommended that the substances selected for phase II include:

- Substances tested during the validation of the rodent tests (pubertal development and thyroid function assays in intact juvenile male and female rats, 15 days intact male rat assay).
- Substances with a mode of action not represented in phase I.
- Substances inert on the thyroid axis.

During the VMGeco meeting in 2015, Linuron was proposed and approved for phase II as it was shown to be active on the thyroid axis in the rodent tests. NH3, an active compound with a mode of action not

represented in phase I (antagonism of thyroid receptor) was selected for phase II. E2 was included as a substance identified inactive on the thyroid axis during the validation of the Amphibian Metamorphosis Assays (TG231). As phase II was carried out by laboratories that did not participate in phase I using new spectrofluorimeter models and a new test medium, the VMGeco advised to include a substance already tested in phase I: PTU. During this meeting the VMGeco agreed that additional data on substances inert on the thyroid axis must be generated.

Preliminary experiments have been performed in the lead laboratory using the proposed substances to determine appropriate testing concentrations. In particular, the solubility in the test medium has been assessed as well as the Maximum Tolerated Concentrations (MTC). To determine the MTC, groups of 20 tadpoles were exposed to a range of concentrations for 72h at 21°C and the mortality was assessed. The highest concentration of the concentration range was set to the solubility limit or 100 mg/L. The MTC was defined as the highest tested concentration allowing at least 80% of survival.

Linuron

Linuron is an herbicide. The juvenile male rat test and 15 days adult male test showed that exposure to Linuron led to a decrease in circulating thyroid hormone concentrations (USEPA 2003; O'Connor et al. 2002). As the exact mode of action of Linuron leading to a decrease in circulating thyroid hormone is not known, the expected result of the XETA was difficult to predict. In the juvenile male rat test this compound has been shown to induce a decrease in T4 and TSH concentrations. These two observations suggest that Linuron could act as a thyroid receptor agonist. If Linuron is a thyroid receptor agonist, an increase in fluorescence is expected in the XETA.

The following concentrations were selected for the phase II validation study: 20; 15; 10; 5; 2.5; 1 mg/L.

NH-3

NH3 is a synthetic TR antagonist (Lim et al. 2002). NH3 inhibits the binding of thyroid hormones to their receptor and cofactor recruitment. As NH3 is a specific TR antagonist a decrease of fluorescence is expected in the T3 spiked XETA test.

The following concentrations were selected: 1; 0.5; 0.1; 0.01; 0.001 mg/l.

17β-estradiol

E2 or 17β-estradiol, is a steroid and estrogenic hormone, it is the primary female sex hormone. E2 was identified as inactive using the Amphibian Metamorphosis Assay during the OECD validation of this test guideline. Besides this result, the rationale for choosing E2 as a test chemical is that it is a potent endocrine active compound that acts via signalling pathways not directly related to the thyroid system. Given that many endocrine-disrupting chemicals possess weak or moderate estrogenic activity, the estrogen E2 could be considered as a representative chemical for estrogenic agents. Therefore, testing E2 over a wider concentration range should provide information about the ability of the test to distinguish thyroid related and unrelated mechanisms (*e.g.* estrogenic activity).

The following concentrations, were selected for phase II: 0.08, 0.4, 1.0, 2.0, 10.0 □g/l. This range is based on the range used for the Amphibian Metamorphosis Assay : 0.08, 0.4, 2.0, 10.0 □g/l.

PTU

The aim of testing again PTU is to provide data on the repeatability of the method over time.

An increase in fluorescence was observed by all the participating laboratories in Phase I and was expected to be observed in phase II.

In phase II PTU was tested at the same concentrations tested in phase I: 1; 3; 10; 30 and 100 mg/L.

Additional inert substances

As the XETA is intended to be used as a screening test, ensuring that substances inert on the thyroid axis will not be identified as positive is a concern. During the validation phase II, XETA experiments on several additional inert substances have been carried out by one or two of the participating laboratories.

Testosterone

Testosterone is a steroid, an androgen, and the primary male sex hormone in mammals. Testosterone was selected following a proposition of the VMGeco during the meeting in 2016. As for E2, the rationale for choosing testosterone as a test substance is that is a potent endocrine active compound acting via signalling pathways not directly related to the thyroid system.

4. The following concentrations, were selected for phase II: 0.08, 0.4, 1.0, 2.0, 10.0 µg/l. This range was chosen to match the range use for E2.

Abamectin, Acetone, Isophorone and Methomyl

In 2016 Wegner et al. from the US EPA selected reference chemicals for thyroid bioactivity screening based on thyroid bioactivity in 'Tier 1' screening assays used by the EPA's Endocrine Disruptor Screening Program (Wegner 2016). Abamectine, Acetone, Isophorone and Methomyl were identified as inactive reference chemicals for thyroid activity. Inactive reference chemicals were selected on the basis of their absence of significant effects on thyroid-responsive endpoints in Tier 1 assays (Amphibian metamorphosis assay, Rat pubertal assay). Absence of thyroid related effect in amphibian or rodent studies from several online databases was also a criterion for selecting these reference chemicals.

Abamectine is an insecticide as well as an acaricide and a nematicide. Acetone is the simplest ketone, commonly used as a solvent or a building block in organic chemistry. Isophorone is a ketone used as a solvent and as a precursor to polymers. Methomyl is a carbamate insecticide highly toxic to humans and wildlife.

4.2. Results

4.2.1. Power analysis

Following the completion and analysis of phase I, the experimental design with 3 runs and 20 embryos per run of each of 5 test concentrations and control was maintained for phase II.

To perform the power analysis following phase II, several concentration-response shapes were simulated to capture the variety of trends observed in the validation studies. The XETA experimental design leads to variance components for run, run-by-treatment, and well-nested within run-by-treatment. This is different from the ecotoxicity experimental designs used in most OECD guidelines where runs refer to tanks, pens, or containers nested within each concentration. Analysis of XETA data treating runs incorrectly as nested within treatment has significant effects on the power properties of the tests. Only the Dunnett and Williams tests are readily analysed with the correct variance structure. The power properties discussed here contain results only for those two tests and it is recommended that Williams' test be used when the data are consistent with a monotone concentration-response and Dunnett's test be used otherwise, following a transformation to achieve normality and homogeneity of variance if

necessary. In both cases, it is important that these tests be carried out using the correct variance structure, as treating runs in the more common fashion seriously degrades the power properties of these tests. The simulations assumed test concentrations (treatments) were geometrically spaced with a constant ratio of 2. Phase 2 studies had somewhat different spacing (e.g., 1, 3, 10, 30, 100 ug/L) for some compounds, but the impact on power would be modest, especially with regard to the effect in the highest test concentration.

Finally, the magnitude of variance observed in phases I and II was simulated using the indicated variance structure. For Williams' test, a one-sided test was done, whereas for Dunnett's test, two-sided tests were simulated since that is more appropriate when the data are not consistent with a monotone concentration-response. It should be emphasized that consistency with a monotone concentration-response does not mean the observed concentration-response must be monotone, only that it does not deviates so strongly as to invalidate a test based on monotonicity. Williams' test is designed to smooth the deviations from monotonicity using the pool-the-adjacent-violators (PAVA) algorithm. The guidance provided earlier applies, namely that if three or more of the five treatment groups are amalgamated using PAVA, then concentration-response monotonicity is questionable and Dunnett's test should be used. Expert statistical judgment on a case-by-case basis may override this general guidance.

Four of the concentration-response shapes simulated are given in Figure 13. A fifth, designated ECPB=1, is not shown. It has a shape similar to that of ECPB=5 except the initial decline is even more rapid. The plots show a maximum effect of 35%. The shapes will vary in predictable fashion according to the maximum effect simulated, which ranged from 30% to 5% for each shape. Power properties of the statistical tests are provided in Table 16. It will be observed that Williams' test has 90% power or greater to detect a 15% effect. By interpolation, 80% power should be achieved to detect a 12-13% effect.

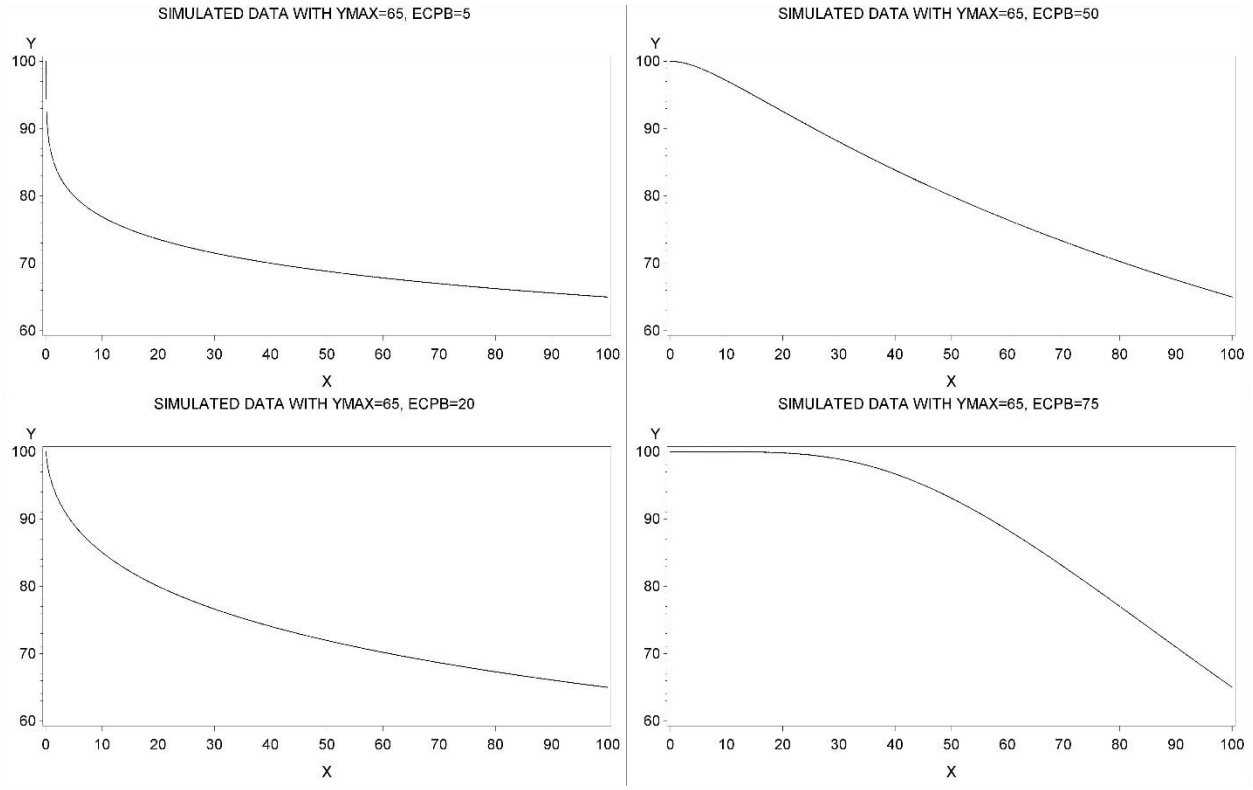

Figure 13 Simulated concentration-response shapes

		Power	
ECPB	%Eff	Dunnett	Williams
1	5	23	36
1	10	50	72
1	15	77	91
1	20	95	99
1	25	100	100
1	30	100	100
5	5	22	34
5	10	51	71
5	15	77	92
5	20	94	99
5	25	99	100
5	30	100	100
25	5	21	36
25	10	51	70
25	15	79	92
25	20	95	98
25	25	99	100
25	30	100	100
50	5	23	35
50	10	51	71
50	15	81	93
50	20	95	99
50	25	99	100
50	30	100	100
75	5	21	34
75	10	53	72
75	15	82	92
75	20	96	99
75	25	100	100
75	30	100	100

Table 16. Power of tests for all concentration-response shape, 20 tadpoles per treatment

It will be seen that Williams' test has 90% or greater power to detect a 15% effect for all concentration-response shapes simulated. Dunnett's test has 90% power to detect a 20% effect and just under 80% power to detect a 15% effect. There is general agreement within the statistical community that 80% power is adequate.

The power simulations are in agreement with the results of phase II. This was confirmed by first finding for each laboratory and chemical the minimum size effect that was found significant by a one-sided test in the direction appropriate for the dataset. Studies where no effect was found significant were excluded. From this subset, the median and mean of these minimum size effects were 14 and 16.4%, respectively, for Dunnett's test and 12 and 12.7%, respectively, for Williams' test. Next, of those laboratory – chemical combinations where no effect was found significant, the maximum size effect observed in those studies was examined. The median and mean of these maximum size effects were 7.5 and 8.3%, respectively, for Dunnett's test and 4% for both mean and median for Williams' test.

An additional power analysis was done following phase II to consider a smaller test design using three concentrations instead of five (Table 17). The five concentration-response shapes simulated describe previously were used. Power properties of the statistical tests are summarized in Table 17. It will be observed that Williams' test has 90% power or greater to detect a 15% effect. By interpolation, 80% power should be achieved to detect a 12-13% effect. That was indeed found to be true in new simulations for the smaller test design.

Additional power analysis was performed following the first WNT commenting round and could be found in Annex 3.

		Power of Williams' Test	
ECPB	%Eff	5 Conc Design	3 Conc Design
1	10	72	
1	12		83
1	15	91	
5	10	71	
5	12		86
5	15	92	
25	10	70	
25	12		83
25	15	92	
50	10	71	
50	12		83
50	15	93	
75	10	72	
75	12		84
75	15	92	

Table 17 Power analysis: key results

4.2.2. False positive rate

The statistical tests used control the false positive results at the 5% level. This is true for 5-concentration and 3-concentration experiments alike. Separate power calculations are done for these two test designs, though the power properties are similar.

4.2.3. Interlaboratory control comparison

The CVs for the same chemical were compared across laboratories. The mean used to calculate the CV was the control mean for the unspiked experiments and the T3 control mean for the spiked experiments. Two methods were used. For the dataset from a single lab, chemical and phase, after the trimming, an ANOVA was done with Rep and Rep*Treatment as random effects and Treatment as the sole fixed effect. The ANOVA removed the treatment effects. The pooled variance of the treatment means from this model were computed, as this is the variance used in statistical tests. Two versions of CV were computed, one using the standard error (square-root of the reported variance) and the other using the standard deviation (square-root of reported variance multiplied by the degrees of freedom). The former is more applicable to the statistical tests, while the latter is more traditional.

In the unspiked mode (Table 18), the mean CVSEs from the four laboratories tended to be relatively close ranging from 7.4 to 10.1. The Belgian lab was the most consistent, as the CVs ranged from 5.4 to 9.5 across the four chemicals tested. This compares to ranges of 5.3 to 13.7 for Portugal (5 chemicals), 7.8 to 13.3 for Japan (4 chemicals), and 5 to 11.2 for France (8 chemicals). No chemical-related pattern was apparent in CVs.

LABORATORY	CHEMICAL	StdErr	DF	Ctrlmean	CVSE	CVSD	mean CVSE	mean CVSD
BELGIUM	E2	171.5	10	3158.7	5.4	17.2	7.42	24.88
	LINURON	228	10	2949.7	7.7	24.4		
	PTU	337.2	10	3533.9	9.5	30.2		
	T3	233.4	16	2887.6	8.1	32.3		
	TESTOSTERONE	259.6	10	4048.9	6.4	20.3		
FRANCE	E2	172.9	10	3379.5	5.1	16.2	7.4	23.85
	LINURON	220	10	3063.1	7.2	22.7		
	NH3	318.4	10	4330.1	7.4	23.3		
	PTU	419.4	10	3499	12	37.9		
	T3	168.3	16	3853.1	4.4	17.5		
	TESTOSTERONE	356.6	10	3185.8	11.2	35.4		
	E2	275	10	2074	13.3	41.9		
JAPAN	NH3	133.4	10	1711.9	7.8	24.6	10.125	31.075
	PTU	224	10	2590.4	8.6	27.3		
	T3	197	8	1824.4	10.8	30.5		
	E2	9	10	83.6	10.8	34		
PORTUGAL	LINURON	13.1	10	95.5	13.7	43.2	8.9	29.18
	NH3	5.3	10	100	5.3	16.7		
	PTU	8.1	10	101.9	8	25.3		
	T3	7.5	16	111.9	6.7	26.7		

Table 18: Distribution of CV (unspiked mode)

. StdErr : standard error, DF : degree of freedom, Ctrl mean : T3 control mean

In spiked mode, the mean CVSEs from the four laboratories ranged from 4.4 to 10.5 the CVs from the Belgium laboratory tended to be the smallest, while those from the Japanese lab tended to be the highest (Table 19). The Belgium lab was the most consistent, with CVs ranging from 2.7 to 6.1 (across 4 chemicals). This compares to Japan with a range from 6.6 to 14.8 (across 3 chemicals), Portugal with a range of 5.9 to 11.6 (across 4 chemicals), and France with a range of 3.8 to 11 (across 7 chemicals). No chemical-related pattern was apparent in CVs.

LABORATORY	CHEMICAL	StdErr	DF	Ctrlmean	CVSE	CVSD	mean CVSE	mean CVSD
BELGIUM	E2T3	117.9	10	4427	2.7	8.4	4.4	14.0
	LINURONT3	263.7	10	4336.2	6.1	19.2		
	PTUT3	297.6	10	5075.4	5.9	18.5		
	TESTOSTERONET3	209.9	10	6503.5	3.2	10.2		
	E2T3	195.6	10	4704.7	4.2	13.1		
	LINURONT3	297.5	10	4968.8	6	18.9		
	NH3T3	229.5	10	6014.9	3.8	12.1		
	PTUT3	551.7	10	5026.4	11	34.7		
	TESTOSTERONET3	499.3	10	4919.7	10.1	32.1		
JAPAN	E2T3	424.6	10	2859.6	14.8	47	10.5	33.2
	NH3T3	258	10	2564.9	10.1	31.8		
	PTUT3	220.6	10	3318.6	6.6	21		
PORTUGAL	E2T3	8.5	10	138.2	6.2	19.5	8.2	26.1
	LINURONT3	16.3	10	140.8	11.6	36.5		
	NH3T3	9.5	10	160.2	5.9	18.8		
	PTUT3	15.2	10	161.6	9.4	29.7		

Table 19: Distribution of CV (spiked mode).

StdErr : standard error, DF : degree of freedom, Ctrl mean : T3 control mean

4.2.4. Interlaboratory test chemical CV comparison

The variability in each experimental group for each chemical was assessed by calculating the coefficient of variation (100*SD/Mean) associated with the mean of the fluorescence values after the 10% trim (*i.e.* 48 fluorescence values in each group, no additional data removal was performed). The table 21 and figure 14 give an overview of the results. The CV for individual experimental groups ranged from 9 to 33%. CV in the Japanese laboratory tends to be higher (mean of all CV: 20%), CV for other laboratories are close: mean of all CV 18% for France, 14% for Belgium, 15% for Japan and Portugal. The CVs are globally better in phase II than in phase I. This is probably due to the changes in the protocol that occurred between phase I and II (recording of empty well position corresponding to dead tadpoles and their removal from the data before statistical analysis).

			-T3						+T3						Mean of CV
			Control	C1	C2	C3	C4	C5	T3	C1	C2	C3	C4	C5	
PTU	France	Mean	3499	3708	3524	3706	3932	3977	5007	5206	5257	5241	5869	6000	
		CV	20%	18%	20%	25%	25%	22%	18%	20%	21%	18%	17%	19%	20%
	Belgium	Mean	3534	3572	3682	3582	3746	4021	5083	4830	5158	5242	5367	5569	
		CV	16%	15%	17%	19%	22%	18%	14%	11%	15%	13%	14%	14%	16%
	Japan	Mean	2590	2792	2746	2421	2692	2782	3319	3207	3304	3444	3477	4183	
		CV	27%	22%	25%	27%	25%	22%	25%	14%	16%	15%	21%	18%	21%
	Portugal	Mean	101,9	103,2	103	102,1	103,7	112,9	161,3	146,9	182,1	178,7	188	194,9	
		CV	18%	18%	18%	13%	12%	10%	20%	9%	15%	14%	13%	14%	15%
NH3	France	Mean	4330	4234	4385	4365	4032	4157	6015	5891	5956	6004	4925	4947	
		CV	17%	16%	15%	17%	13%	13%	11%	11%	10%	11%	15%	12%	13%
	Japan	Mean	1701	1950	1823	1701	1663	1645	2650	2965	2710	2576	2176	2005	
		CV	32%	22%	26%	19%	18%	21%	28%	26%	24%	23%	21%	17%	23%
	Portugal	Mean	102,3	103,1	98,93	106,9	104,3	103,9	162,8	157,2	158	142	122,7	116,7	
		CV	14%	13%	14%	16%	11%	10%	16%	16%	11%	15%	12%	11%	13%
	France	Mean	3063	3595	3732	3686	3557	3798	4944	5514	5295	5307	4881	5253	
		CV	14%	16%	18%	14%	15%	14%	16%	13%	11%	19%	13%	11%	14%
Linuron	Belgium	Mean	2950	2892	2876	3316	3332	3592	4336	4077	4257	4315	4275	4345	
		CV	14%	12%	16%	16%	15%	19%	12%	17%	14%	17%	14%	13%	15%
	Portugal	Mean	95,51	93,09	92,67	96,9	107,4	107,4	140,8	134,6	141,8	139	134,3	134,1	
		CV	24%	21%	22%	23%	20%	23%	14%	14%	19%	28%	22%	18%	21%
	France	Mean	3380	3375	3360	3654	3284	3218	4698	4585	4549	4695	4651	4924	
		CV	18%	15%	16%	16%	20%	15%	12%	11%	14%	12%	14%	12%	15%
E2	Belgium	Mean	3154	3081	3127	3259	3114	3118	4426	4609	3997	4215	4252	4495	
		CV	15%	14%	17%	18%	15%	17%	12%	13%	12%	13%	12%	14%	14%
	Japan	Mean	2074	2146	2023	2051	2234	2082	2860	3024	3192	3157	3050	2875	
		CV	29%	27%	25%	24%	25%	25%	25%	26%	25%	22%	30%	20%	25%
	Portugal	Mean	84,11	82,55	80,86	84,28	80,61	79,84	137,2	135,1	141,8	130,7	130,8	125,3	
		CV	20%	19%	18%	20%	16%	20%	13%	10%	10%	15%	15%	12%	16%
Abamectine	France	Mean	3851	3939	3986	3949	3766	3899	5476	5429	5354	5159	5254	5437	
		CV	16%	15%	14%	13%	14%	17%	13%	17%	16%	12%	14%	13%	15%
Acetone	France	Mean	3256	3478	3471	3609	3297	3350	5197	5016	4755	5051	4811	4918	
		CV	25%	18%	23%	24%	24%	20%	21%	21%	21%	23%	23%	18%	22%
Isophorone	France	Mean	3774	3810	3820	3984	3840	3913	5585	5895	5178	5387	5175	6030	
		CV	28%	31%	31%	31%	30%	29%	28%	33%	26%	28%	25%	30%	29%
Metholmyl	France	Mean	4083	4370	4214	4281	3944	4164	5515	5672	5362	5453	5075	5341	
		CV	14%	15%	19%	16%	17%	12%	9%	14%	13%	13%	16%	13%	14%
Testosterone	France	Mean	3209	3443	2954	2933	3004	3150	4920	4566	4629	4765	4795	4212	
		CV	22%	21%	21%	22%	20%	28%	18%	16%	15%	21%	20%	18%	20%
	Belgium	Mean	4043	4361	4053	4185	3737	4139	6498	6065	6013	6584	6060	6366	
		CV	21%	17%	14%	14%	15%	13%	13%	15%	13%	12%	11%	13%	14%

Table 20: Overview of the phase II mean and CV

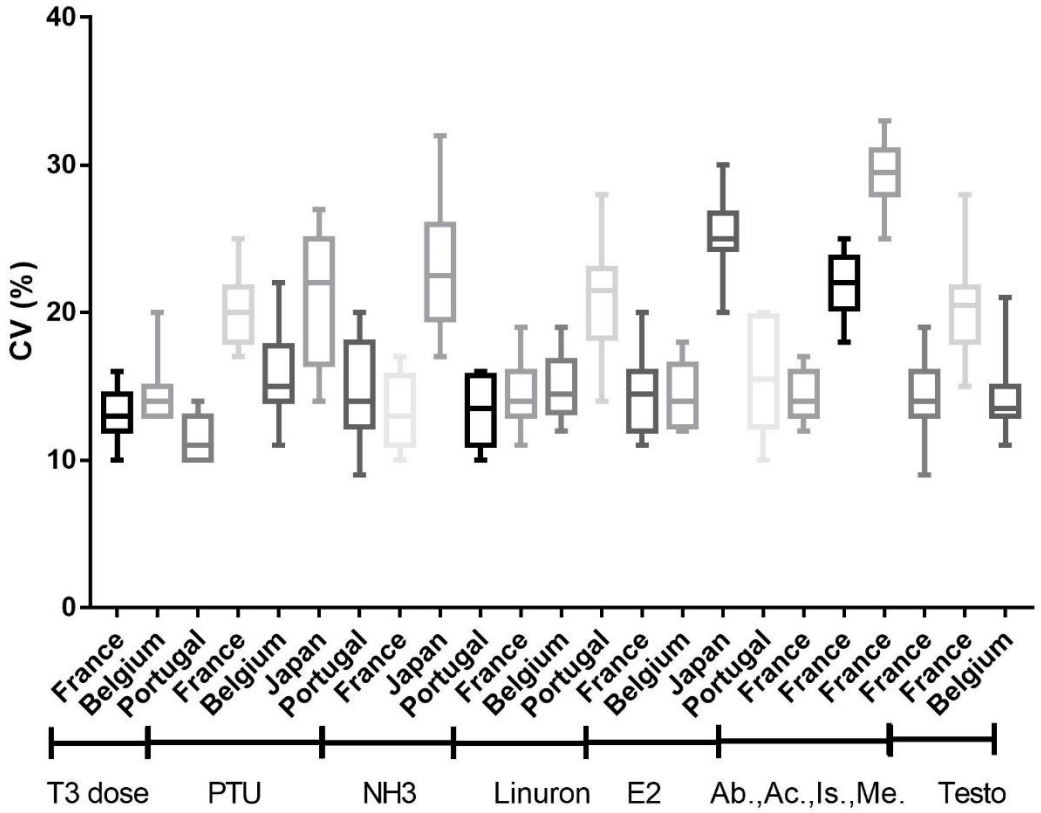

Figure 14: Overview of the phase II CV in each experimental group

4.2.5. Calibration (T3 concentration response)

Three runs of a concentration response of eight concentrations of T3 were performed. The concentration response T3 data is included in figure 15 below showing the mean and SEM of the fluorescence for each concentration of T3 in each laboratory.

The increase in fluorescence in each laboratory is clearly obtained with increasing concentrations of T3. The three laboratories obtained consistent increases in fluorescence over 12% at every concentration of T3 above 0.65 µg/L. For these three laboratories, the statistical analysis shows all increases in fluorescence above 0.65 µg/L to be significant. In addition the Portuguese laboratory founds the increase obtained at 0.65 µg/L to be significant (Table 21). During the training phase, Japan provided one non-conforming experiment (T3 control induction lower than 20%), due to time constraints Japan did not repeat this experiment. However, experiments carried out after this experiment in the Japanese

laboratory satisfied this validity criteria suggesting that the problem have been caused by a lack of experience in performing the experiment.

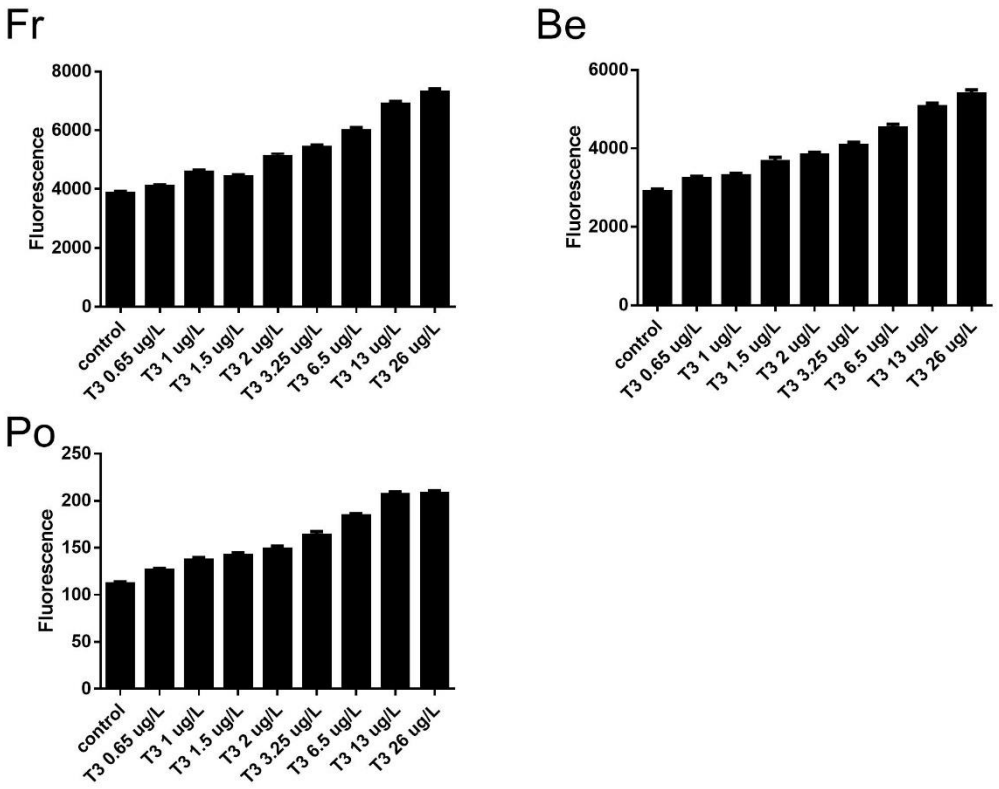

Figure 15 : Mean and SEM of fluorescence intensities in the T3 concentration response calibration experiment

Country	T3 (µg/l)							
	0,65	1	1,5	2	3,25	6,5	13	26
France	6	19	15	32	40	55	79	89
Belgium	11	14	27	32	41	56	75	86
Portugal	13	23	27	33	46	64	85	86

Table 21: Percentage of fluorescence variations in the T3 experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

4.2.6. PTU

Figure 16 below shows the mean and SEM for each concentration of PTU in each laboratory. A consistent increase in fluorescence in each laboratory was obtained with increasing concentrations of PTU in the presence of T3 and at the maximum concentration in the absence of T3 for two laboratories.

The statistical analysis of the data shows a significant increase equal or greater to 12% at 30 and 100 mg/L for France and at 100 mg/L for Belgium in unspiked mode. No active concentrations were detected for Portugal and Japan with non-significant increase of 7 and 10% observed at 100 mg/l.

In spiked mode, a significant increase is detected for concentrations above 1 mg/l for Portugal, 10 mg/l for France and 30 mg/l for Belgium and Japan. All these fluorescence increases were equal or greater to 12% except for Belgium (10%) (Table 22).

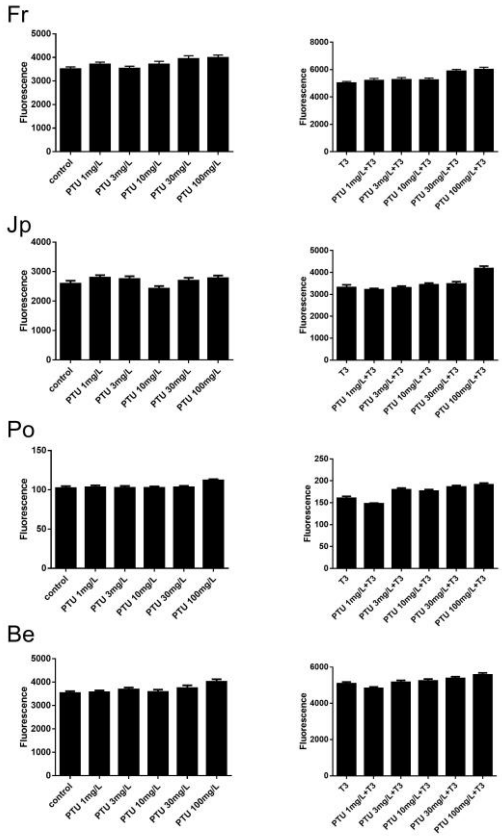

Figure 16: Mean and SEM of fluorescence intensities in the PTU T3 concentration response experiment.

Country	PTU (mg/l)									
	-T3					+T3				
	1	3	10	30	100	1	3	10	30	100
France	6	1	6	12	14	4	5	5	17	20
Japan	8	6	-7	4	7	-3	0	4	5	26
Portugal	1	0	0	1	10	-8	12	10	16	19
Belgium	1	4	1	6	14	-5	1	3	6	10

Table 22: Percentage of fluorescence variations in the PTU experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

4.2.7. Linuron

Figure 12 below shows the mean and SEM for each concentration of Linuron for three laboratories. The Japanese laboratory didn't provide results for Linuron due to time constraints. An increase in fluorescence in each laboratory was obtained with increasing concentrations of Linuron in the absence of T3 and no effects are observed in the presence of T3. Increases of 12% were observed but not found to be statistically significant in the Portuguese test.

The statistical analysis of the data shows these fluorescence increases to be significant and over 12% for concentrations over 5 mg/ for Belgium, and for all concentrations for France.

In spiked mode, no significant fluorescence variations were detected for the three laboratories (Table 17). An increase of 12% was observed but not found to be statistically significant in the French test.

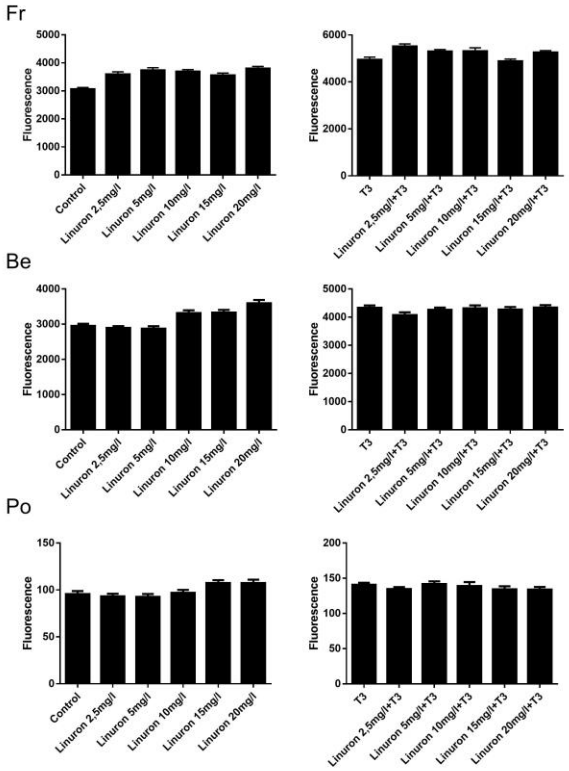

Figure 17: Mean and SEM of fluorescence intensities in the linuron experiment

Country	Linuron (mg/l)									
	-T3					+T3				
	2,5	5	10	15	20	2,5	5	10	15	20
France	17	22	20	16	24	12	7	7	-1	6
Belgium	-2	-3	12	13	22	-6	-2	0	-1	0
Portugal	-3	-3	1	12	12	-4	1	-1	-5	-5

Table 23: Percentage of fluorescence variations in the linuron experiment

The results corresponding to a statistically significant variation of fluorescence are highlighted in green

4.2.8. NH3

Figure 18 below shows the mean and SEM for each concentration of NH3 for three laboratories. The Belgium laboratory didn't provide results for NH3 due to time constraints. A decrease in fluorescence in each laboratory was obtained with increasing concentrations of NH3 in the presence of T3 and flat concentration responses are observed in the absence of T3.

The statistical analysis of the data shows these fluorescence decreases to be significant and over 12% for concentrations over 0.01 mg/l for Portugal and over 0.1 mg/L for France and Japan. A decrease of 12% at the lowest concentration in the Japanese test wasn't found significant.

In unspiked mode, no significant fluorescence variations were detected for the three laboratories.

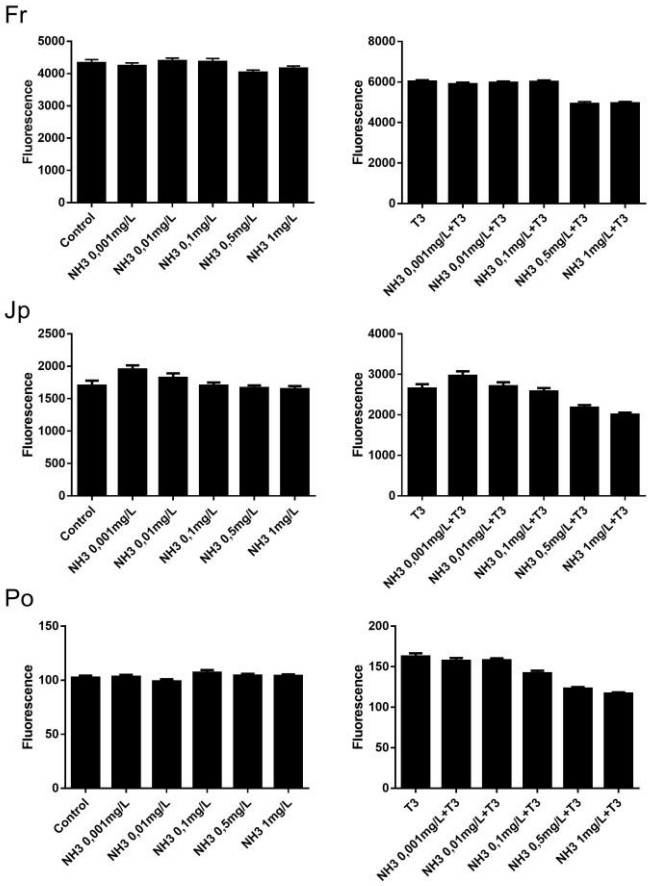

Figure 18: Mean and SEM of fluorescence intensities in the NH3 concentration response experiment

Country	NH3 (mg/l)									
	-T3					+T3				
	0,001	0,01	0,1	0,5	1	0,001	0,01	0,1	0,5	1
France	-2	1	1	-7	-4	-2	-1	0	-18	-18
Japan	15	7	0	-2	-3	12	2	-3	-18	-24
Portugal	1	-3	4	2	2	-3	-3	-13	-25	-28

Table 24: Percentage of fluorescence variations in the NH3 experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

4.2.9. E2

Figure 19 below shows the mean and SEM for each concentration of E2 for the participating laboratories. Flat concentration responses with erratic fluorescence variations could be observed both in spiked and unspiked modes.

The statistical analysis of the data shows fluorescence variations to be statistically significant for concentrations over 1 µg/L for Portugal but

the fluorescence variations remained under 10%. In the Japanese test a 12% increase was observed but not found statistically significant.

In unspiked mode, no significant fluorescence variations were detected for the three laboratories (Table 19).

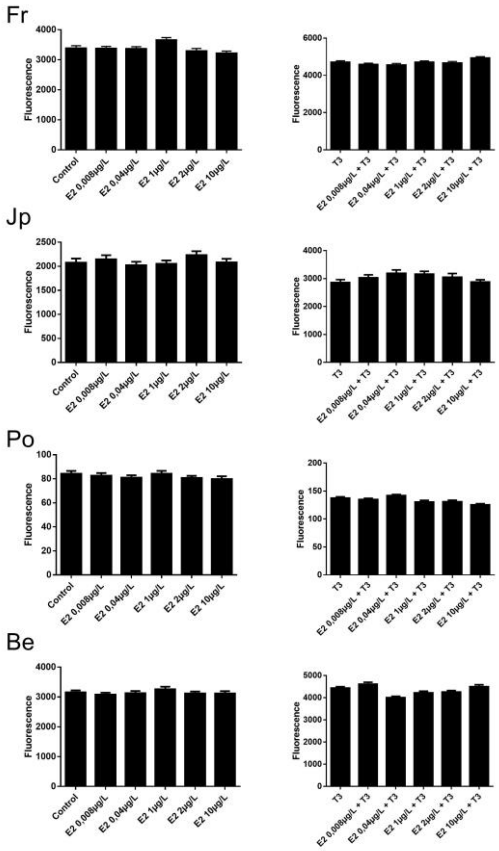

Figure 19: Mean and SEM of fluorescence intensities in the E2 concentration response experiment

Country	E2 (µg/l)									
	-T3					+T3				
	0,008	0,04	1	2	10	0,008	0,04	1	2	10
France	0	-1	8	-3	-5	-2	-3	0	-1	5
Japan	3	-2	-1	8	0	6	12	10	7	1
Portugal	-2	-4	0	-4	-5	-2	3	-5	-5	-9
Belgium	-2	-1	3	-1	-1	4	-10	-5	-4	2

Table 25: Percentage of fluorescence variations in the E2 experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

4.2.10. Testosterone

6. France and Belgium tested testosterone. Figure 15 below shows the mean and SEM for each concentration of Testosterone. Flat

concentration responses with erratic fluorescence variations lower than 10% could be observed both in spiked and unspiked modes.

The statistical analysis of the data shows some fluorescence variations to be significant. In spiked mode the concentration of 10 mg/L was statistically significant for France and in unspiked mode 10 mg/L was statistically significant for Belgium. A 14% fluorescence decrease was observed at 10 mg/ for France (Table 20).

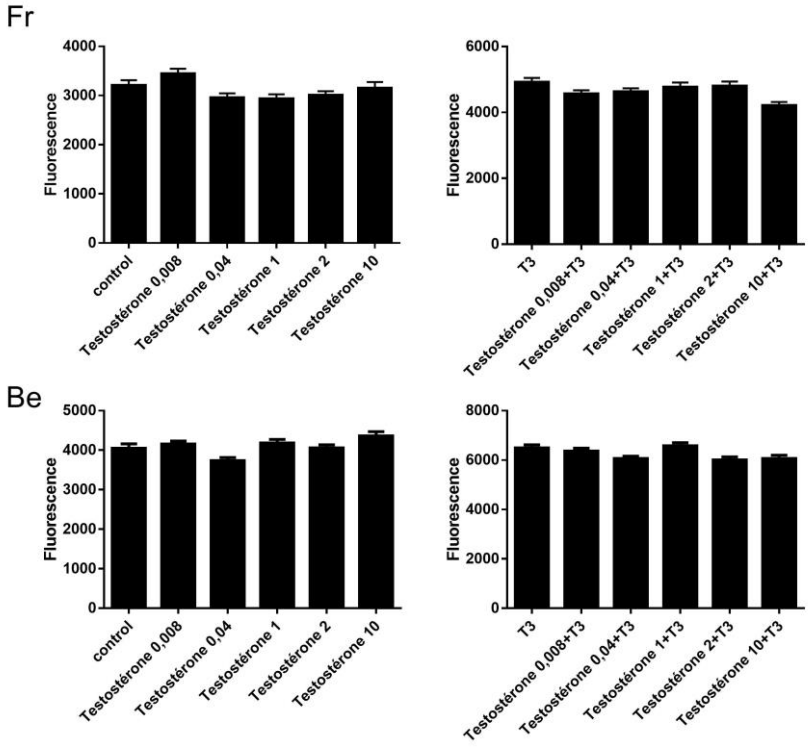

Figure 20: Mean and SEM of fluorescence intensities in the testosterone concentration response experiment

Country	Testosterone (mg/l)									
	-T3					+T3				
	0,008	0,04	1	2	10	0,008	0,04	1	2	10
France	7	-8	-9	-6	-2	-7	-6	-3	-3	-14
Belgium	3	-8	4	0	8	-2	-7	1	-7	-7

Table 26: Percentage of fluorescence variations in the Testosterone experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

4.2.11. Abamectin, Acetone, Isophorone, Methomyl

7. All these four substances were tested by France to produce additional data on substances inactive on the thyroid axis. Figure 21

below shows the mean and SEM for each concentration of these four substances. Erratic fluorescence variations lower than 12% could be observed both in spiked and unspiked modes for all substances. None of these variations in fluorescence were statistically significant except for Isophorone at 6 mg/L which is significant in unspiked mode (9% increase) (Table 27).

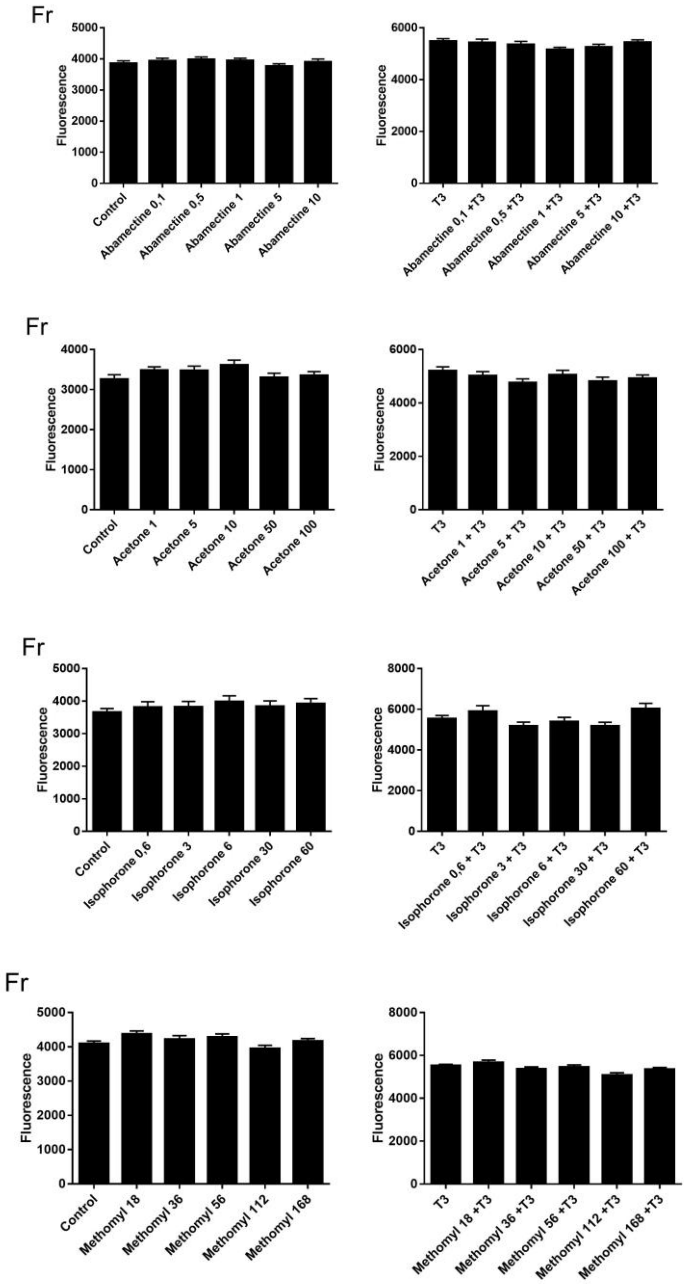

Figure 21: Mean and SEM of fluorescence intensities in the reference inert substances experiments

Country	Abamactine (mg/l)									
	-T3					+T3				
	0,1	0,5	1	5	10	0,1	0,5	1	5	10
France	2	4	3	-2	1	-1	-2	-6	-4	-1

Country	Acetone (mg/l)									
	-T3					+T3				
	1	5	10	50	100	1	5	10	50	100
France	7	7	11	1	3	-3	-9	-3	-7	-5

Country	Isophorone (mg/l)									
	-T3					+T3				
	0,6	3	6	30	60	0,6	3	6	30	60
France	4	4	9	5	7	7	-6	-3	-6	9

Country	Metholmyl (mg/l)									
	-T3					+T3				
	18	36	56	112	168	18	36	56	112	168
France	7	3	5	-3	2	3	-3	-1	-8	-3

Table 27: Percentage of fluorescence variations in the reference inert substance experiment.

The results corresponding to a statistically significant variation of fluorescence are highlighted in green.

4.3. Chemical Analyses

Selected samples were retained and frozen for quantitative chemistry to ensure that the laboratories were able to accurately prepare the chemicals used in the assay at the correct concentrations. Chemical analysis was performed on samples from Japan, France and Portugal for experiments testing PTU, Linuron, NH3 and E2.

Participating laboratories were asked for four substances to:

-sample exposure solutions at the highest and lowest concentration before starting the experiments at D0

-sample exposure solutions at the highest and lowest concentration after 24 h of exposure. The decision to sample after 24h of exposure was taken after the start of the of the ring test and the completion of the first experiments, therefore some of these samples were not provided by the participating laboratories.

All solutions were stored at -20° and sent to the Japanese CRO “IDEA consulting” by the participating laboratories. IDEA reported that all “before exposure” samples were received and analysed. Concerning the “24h samples”, four samples from the Portugese laboratory were missing and one could not be identified unambiguously because of unreadable information written on the tube. IDEA measured the concentrations of the test compounds using LC-MS. The limits of quantification using this method are detailed in the following table.

Chemical	Limit of quantification (ng/ml)
PTU	17
Linuron	13
NH3	0.031
E2	0.087

Table 28: Quantification limits for the chemical analysis

Figure 22 and 23 give an overview of the results showing the measured concentrations as percentages of the nominal concentrations for the different laboratories and chemicals.

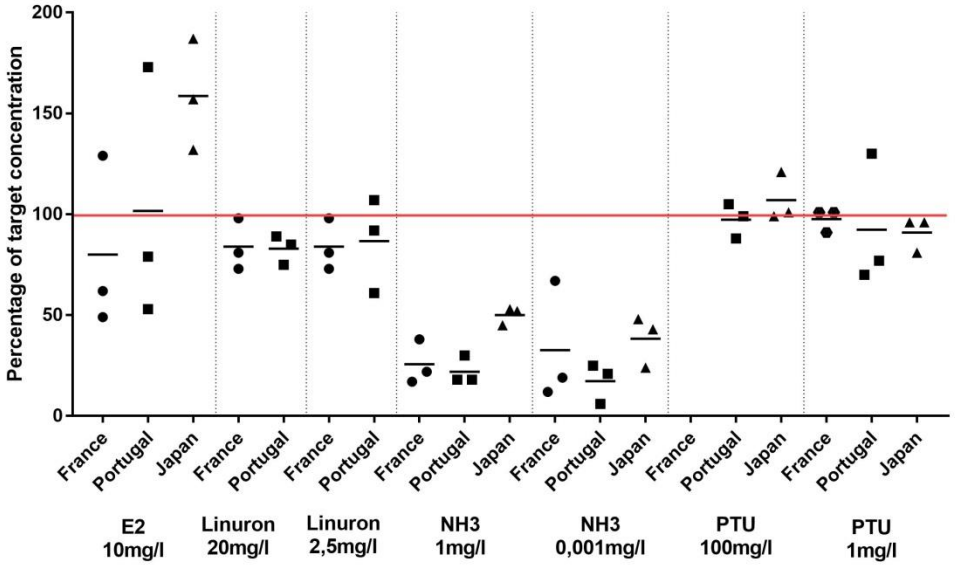

Figure 22. Percentage of target concentrations reached in the exposure medium for each rune of XETA experiments before the exposure.

Symbols indicate the percentage of target concentration for each individual runs. Lines indicate the mean of the three runs values.

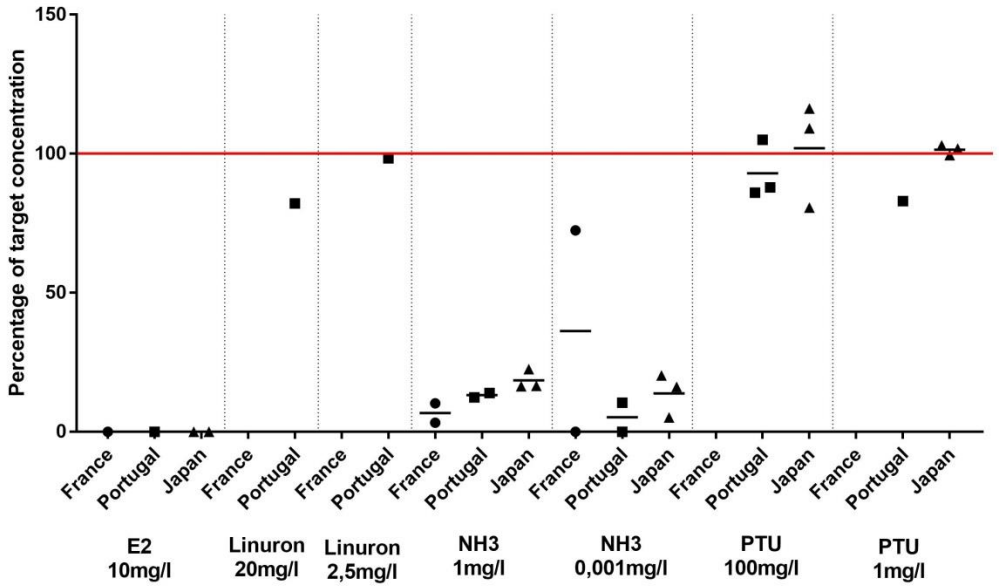

Figure 23. Percentage of target concentrations reached in the exposure medium for each run of XETA experiments after 24h of exposure.

Symbols indicate the percentage of target concentration for each individual run. Lines indicate the mean of the three runs values.

4.3.1. E2

The lowest nominal concentration tested (0.08 µg/L) was below the limit of detection (0.087 µg/L). Samples prepared at the lowest tested concentration were therefore not analysed. Overall Table 29 and Figure 22 show that the measured concentrations represent between 62 and 187 % of the nominal concentration. The mean of the measured concentrations for the 3 runs is close to 100% of the nominal concentration for the Portuguese and French samples and close to 160% for the Japanese samples. After 24 h the E2 concentration was under the limit of detection for the samples (Table 22). This could indicate that E2 was completely absorbed by the tadpoles as E2 is very stable in water solution.

Country	Run	Nominal E2 Concentration (µg/L)	Measured E2 Concentration (µg/L)	Measured/Nominal (%)
France	R1	10	12.9	129
	R2	10	6.2	62
	R3	10	4.9	49
Portugal	R1	10	7.9	79
	R2	10	5.3	53
	R3	10	17.3	173
Japan	R1	10	18.7	187
	R2	10	15.7	157
	R3	10	13.2	132

Table 29 : Nominal and measured concentrations for E2 before exposure

Country	Run	Nominal E2 Concentration (µg/L)	Measured E2 Concentration (µg/L)	Measured/Nominal (%)
France	R1	10	＜0.087	-
Portugal	R3	10	＜0.087	-
Japan	R1	10	＜0.087	-
	R2	10	＜0.087	-
	R3	10	＜0.087	-

Table 30 : Nominal and measured concentrations for E2 after 24 h of exposure

4.3.2. Linuron

Overall table 31 and figure 22 show that the measured concentrations represent between 61 and 107 % of the nominal concentrations. The mean measured concentration for the 3 runs is close to 85% of nominal for the Portuguese and French samples. Comparable measured concentrations are found after 24 h indicating that Linuron is stable in the test medium and is not metabolized to a high degree by the tadpoles.

Country	Runs	Nominal Linuron Concentration (mg/L)	Measured Linuron Concentration (mg/L)	Measured/Nominal (%)
France	R1	1	1.0	98
	R2	1	0.8	81
	R3	1	0.7	73
Portugal	R1	20	17.7	89
	R2	20	17.0	85
	R3	20	15.1	75
Portugal	R1	2.5	2.3	92
	R2	2.5	2.7	107
	R3	2.5	1.5	61

Table 31: Nominal and measured concentrations for Linuron before exposure

Country	Run	Nominal Linuron Concentration (mg/L)	Measured Linuron Concentration (mg/L)	Measured/Nominal (%)
Portugal	R2	2.5	2.5	98
	R2	20	16	82

Table 32 : Nominal and measured concentrations for linuron after 24 h of exposure

4.3.3. NH3

Overall table 33 and figure 22 show that the measured concentrations represent between 6 and 67 % of the nominal concentration. The mean measured concentration for the 3 runs was under 50% of nominal for the three participating laboratories. As limited data are available concerning NH3 solubility the solubilization of NH3 was possibly incomplete.

Country	Run	Nominal NH3 Concentration (mg/L)	Measured NH3 Concentration (mg/L)	Measured/Nominal (%)
France	R1	1	0.2	17
	R2	1	0.2	22
	R3	1	0.4	38
France	R1	0.001	0.00012	12
	R2	0.001	0.00019	19

	R3	0.001	0.00067	67
Portugal	R1	1	0.2	18
	R2	1	0.3	30
	R3	1	0.2	18
Portugal	R1	0.001	0.00021	21
	R2	0.001	0.00025	25
	R3	0.001	0.00006	6
Japan	R1	1	0.4	45
	R2	1	0.5	52
	R3	1	0.5	53
Japan	R1	0.001	0.00024	24
	R2	0.001	0.00048	48
	R3	0.001	0.00043	43

Table 33 : Nominal and measured concentrations for NH3 before exposure

Country	**Run**	**Nominal NH3 Concentration (mg/L)**	**Measured NH3 Concentration (mg/L)**	**Measured/Nominal (%)**
France	R2	1	0.033	3
	R3	1	0.10	10
France	R1	0.001	<0.000031	-
	R2	0.001	0.00072	72
Portugal	R2	1	0.12	12
	R3	1	0.14	14
Portugal	R2	0.001	0.00011	11
	R3	0.001	<0.000031	-
Japan	R1	1	0.17	17
	R2	1	0.16	16
	R3	1	0.23	23
Japan	R1	0.001	0.000052	5
	R2	0.001	0.00020	20
	R3	0.001	0.00016	16

Table 34: Nominal and measured concentrations for NH3 after 24 h of exposure

4.3.4. PTU

Overall table 35 and figure 22 show that the measured concentrations represent between 70 and 121 % of the nominal concentration. The mean measured concentration for the 3 runs was close to 100% for the

three laboratories. Comparable measured concentrations after 24h of exposure indicate that PTU is stable in the test medium and not metabolized to a high degree by the tadpoles.

Country	Run	Nominal PTU Concentration (mg/L)	Measured PTU Concentration (mg/L)	Measured/Nominal (%)
France	R1	1	1.0	101
	R2	1	1.0	101
	R3	1	0.9	91
Portugal	R1	100	99.2	99
	R2	100	88.0	88
	R3	100	105.4	105
Portugal	R1	1	1.3	130
	R2	1	0.7	70
	R3	1	0.8	77
Japan	R1	1	1.0	96
	R2	1	0.8	81
	R3	1	1.0	96
Japan	R1	100	99.0	99
	R2	100	121.1	121
	R3	100	101.4	101

Table 35: Nominal and measured concentrations for PTU before exposure

Country	Replicat	Nominal PTU Concentration (mg/L)	Measured PTU Concentration (mg/L)	Mesured/Nominal (%)
Japan	R1	100	109	109
	R4	100	81	81
	R5	100	116	116
Portugal	R1	100	88	88
	R2	100	86	86
	R3	100	105	105
Portugal	R3	1	0.83	83
Japan	R1	1	1.0	100
	R4	1	1.0	103
	R5	1	1.0	102

Table 36 : Nominal and measured concentrations for PTU after 24 h of exposure

4.4. Discussion

4.4.1. Control groups

All differences between medium only & T3 and T3 & T4 controls were found statistically significant in every experiment.

The T4 control group induced levels that never reached the saturation level of the spectrofluorimeters with the applied settings showing that every laboratory could effectively quantify the highest levels of fluorescence reached by the tadpoles.

Criteria of validity for the XETA included reaching a 20% increase in fluorescence in the T3 control group. Inductions ranged from 35 to 61% for France, 28 to 63% for Japan. 46 to 63% for Portugal and 40 to 61% for Belgium (Figure 24).

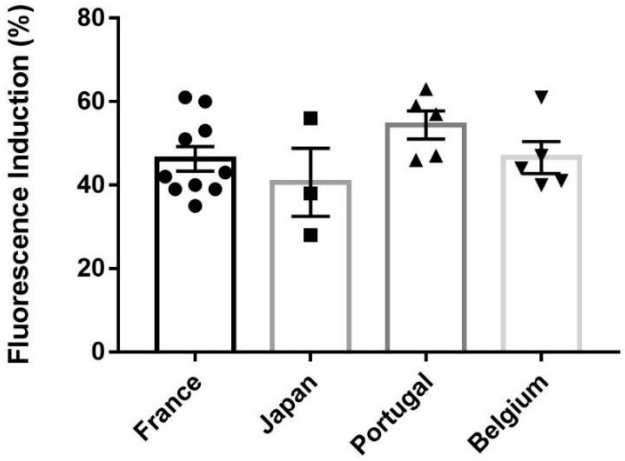

Figure 24 : Percentage of fluorescence induction in the T3 control groups during phase II

4.4.2. Calibration (T3 concentration response)

T3 was identified as a thyroid active molecule with a LOEC of 0.65 µg/L for Portugal, 1 µg/L for France and Belgium. Japan didn't provide a valid experiment for T3 (please see section 4.2.5). This result shows that overall the calibration experiments worked as intended for three

laboratories that were all able to obtain a similar sensitivity in response to T3.

4.4.3. PTU

As for phase I PTU gave a clear increase in fluorescence across all laboratories at the higher concentrations. A response could be observed in the spiked groups above 10 mg/L and in the unspiked groups at the maximum concentration (100 mg/L). During phase II, all laboratories identified PTU as a thyroid active molecule with LOEC of 3 mg/L for Portugal, 30 mg/L for France and 100 mg/L for Belgium and Japan.

4.4.4. Linuron

Linuron induced an increase in fluorescence across all laboratories specifically in the unspiked mode. France and Belgium identified Linuron as a thyroid active molecule with a LOEC of 2.5 mg/L for France and 10 mg/L for Belgium. Despite a fluorescence increase of 12% for the two highest concentrations, William's test found the fluorescence variation in the Portuguese experiment to be not significant. The observed increase suggests that Linuron acts as a weak agonist of the TR, this is in accordance with results from the male juvenile rats test where this compound has been shown to induce a decrease in T4 and TSH concentrations (USEPA 2003, 2004; O'Connor et al. 2002). Fluorescence induces by TR agonists in spiked mode are lower than in unspiked mode, for example at the highest tested concentration TRIAC induced over 80% in unspiked mode and 16 % in spiked mode. As Linuron induced considerably lowest inductions (20%) in unspiked mode than TRIAC, it makes sense that its effects in spiked mode are not detectable. The low fluorescence induction shows that if Linuron is a TR agonist it is a considerably weaker agonist than T4 or TRIAC. The limited concentration response observed for Linuron might be due to its low potential to activate the receptor leading to reach a maximal observable effect of 20%.

4.4.5. NH3

As expected for an antagonist of the thyroid receptor, NH3 induces a decrease in fluorescence across all laboratories specifically in spiked mode. All laboratories identified NH3 as a thyroid active molecule with LOEC of 0.1 mg/L for Portugal, 0.5 mg/L for France and Japan. No significant fluorescence decrease is observed in unspiked mode as expected considering that the embryo is not synthetizing TH at this developmental stage.

Considering that the measured concentrations for NH3 are lower than 30% less than the nominal concentrations these LOEC are overestimated.

4.4.6. E2

E2 was identified as a thyroid inactive molecule by the four participating laboratories. This is in accordance with the result for E2 of the Amphibian Metamorphosis Assay showing E2 to be thyroid inactive in Xenopus tadpoles.

4.4.7. Testosterone

Testosterone was found to be inactive by the Belgium laboratory, however the French laboratory identified the highest concentration (10 mg/L) to be active in spiked mode with a reduction in fluorescence of 14%. The maximal fluorescence variation in the Belgium experiment was 8% and none of the tested concentrations gave statistically different variations from control values.

The expected activity of testosterone on the Xenopus thyroid axis at this developmental stage is unclear. To date no Amphibian Metamorphosis Assay or LAGDA results are available for Testosterone but Gray et al. 1990 reported that treating tadpoles for 6 days from stage 52 with 3.4 µM (1 mg/L) testosterone completely blocked the reduction in body weight, the shrinkage of the head and alimentary canal induced by 1 nM T3 (0.65 µg/L). The authors mentioned that this result is consistent with the findings of Frieden and Naile 1955 and Roth et al. 1941, 1942 and 1943 who reported that testosterone inhibited T3 induced metamorphosis in other amphibians (*B. bufo japonicus* and *R. temporuriua*).

The decrease of fluorescence observed by the French laboratory in spiked mode is in accordance with the effect of testosterone describes in Gray et al. 1990.

Considering that only the highest tested concentration of testosterone induced a statistically significant decrease in fluorescence, which was limited to 14%, one possibility is that testosterone could exhibit weak anti thyroid activity at supra-physiological concentrations (10 mg/L) that was not seen by the Belgium laboratory because of a slightly lower sensitivity in this experiment or a difference in the effective exposure concentration. In the absence of chemical analysis results to confirm the exposure concentration is not possible to conclude on the effect of testosterone in the XETA.

Additional experiments using a concentration range of testosterone, including concentrations above and below 10 mg/L, and proper chemical analysis to confirm the exposure concentrations, needs to be performed to provide more detailed information on the effect of testosterone in the XETA.

4.4.8. Reference thyroid inactive molecules

The four selected reference thyroid inactive molecules were found inactive in the XETA. The result of the XETA is therefore in accordance with the results obtained for these molecules by the Amphibian Metamorphosis Assay , the male juvenile rat test and female juvenile rat test (Wegner et al. 2016).

4.5. Conclusions

The validation effort for the XETA allowed to obtain a definitive protocol and to define suitable data treatment, statistical approach and a decision logic to classify the test substance into thyroid active or inactive categories. The protocol was successfully transferred to five laboratories from different OECD countries. After having performed a series of calibration experiments to find the correct settings for the spectrofluorimeter and acquire the proficiency for the assay, the different laboratories obtained the expected results for the chemicals chosen based on the decision tree except the identification of Linuron as thyroid inactive by the Portugal laboratory (Table 37).

The result obtained with Testosterone need additional experiments to be explained as the effect of high concentrations of testosterone on the thyroid system of Xenopus tadpoles is unclear. The XETA validation results demonstrate that the assay provides reasonable sensitivity with the chemicals tested and is reproducible across laboratories.

Despite the difference in effect levels (table 38), all laboratories accurately classified the tested reference substances as active or inactive, with the exception of Linuron for Portugal. Several parameters should have affected the effect levels. Different fluorescence readers were used, typically each laboratory used a fluorescent reader available in house, none of the participating laboratories bought a specific reader. The lowest LOEC for T3, PTU and NH3 were obtained by the Portuguese laboratory in phase II and the US laboratory in phase I obtained the lowest LOEC for PTU and TRIAC. Both laboratories used a Synegy 2 model. Another difference arises from the time each laboratory could have spent for finding the best settings for the spectrofluorometer and for training before testing the substances. Due to time constraints, the Japanese laboratory during phase II has shortened the training phase including the optimisation of the spectrofluorimeter settings and obtained the highest LOEC.

Chemical	Expected classification	Laboratory					
		France	Japan lab1	USA	Japan lab2	Belgium	Portugal
T3	Thyroid active	Thyroid active	Thyroid active	Thyroid active		Thyroid active	Thyroid active
PTU	Thyroid active	Thyroid active	Thyroid active	Thyroid active	Thyroid active	Thyroid active	Thyroid active
T4	Thyroid active	Thyroid active	Thyroid active	Thyroid active			
TRIAC	Thyroid active	Thyroid active	Thyroid active	Thyroid active			
Cefuroxime	Thyroid inactive	Thyroid inactive	Thyroid inactive	Thyroid inactive			
Linuron	Thyroid active	Thyroid active				Thyroid active	Thyroid inactive
NH3	Thyroid active	Thyroid active			Thyroid active		Thyroid active
Testosterone	Unclear	Thyroid active			Thyroid inactive		
E2	Thyroid inactive	Thyroid inactive			Thyroid inactive	Thyroid inactive	Thyroid inactive
Abamectine	Thyroid inactive	Thyroid inactive					
Acetone	Thyroid inactive	Thyroid inactive					
Isophorone	Thyroid inactive	Thyroid inactive					
Metholmyl	Thyroid inactive	Thyroid inactive					

Table 37 : Overview of the ring test results

LOECs	Phase 1			Phase 2			
	France	Japan	USA	Portugal	France	Belgium	Japan
T3 range 1	3.25	3.25	3.25	NT	NT	NT	NT
T3 range 2	NT	NT	NT	**0.65**	1	1	NT
PTU	100	100	**3**	**30**	**30**	100	100
Linuron	NT	NT	NT	-	**2.5**	15	NT
NH3	NT	NT	NT	**0,1**	0.5	NT	0.5
T4	0.1	**0.01**	0.1	NT	NT	NT	NT
TRIAC	0.01	0.01	**0.001**	NT	NT	NT	NT

Table 38 : Overview of the LOECs.

4.6. Complementary elements

4.6.1. Saturation control

We decided in the draft test guideline to use a concentration of 10 mg/L of T4 as a saturation control instead of a cotreatment of 10 mg/L of T4 and 3.25 µg/L of T3. This is supported by results from the phase I of validation showing that the plateau of the T4 concentration response curve starts with concentrations above 1 mg/L and that adding T3 has no effect on the level of fluorescence (see figure 11).

4.6.2. Phenobarbital

During the WNT commenting round, a question was raised to know if the exposure time use for the XETA is sufficient to induce the UDPGT

and therefore detect UDPGT modulators. The lead laboratory performed a XETA using phenobarbital as a test substance. In spiked mode, Phenobarbital is inducing a significant decrease of fluorescence for all the concentrations tested (figure 25). No significant fluorescence variations are detected in unspiked mode. This result is in accordance with the mode of action of phenobarbital known to induce the expression of UDPGT and consequently accelerating the clearance of TH. This shows the XETA to detect UGPGT modulators.

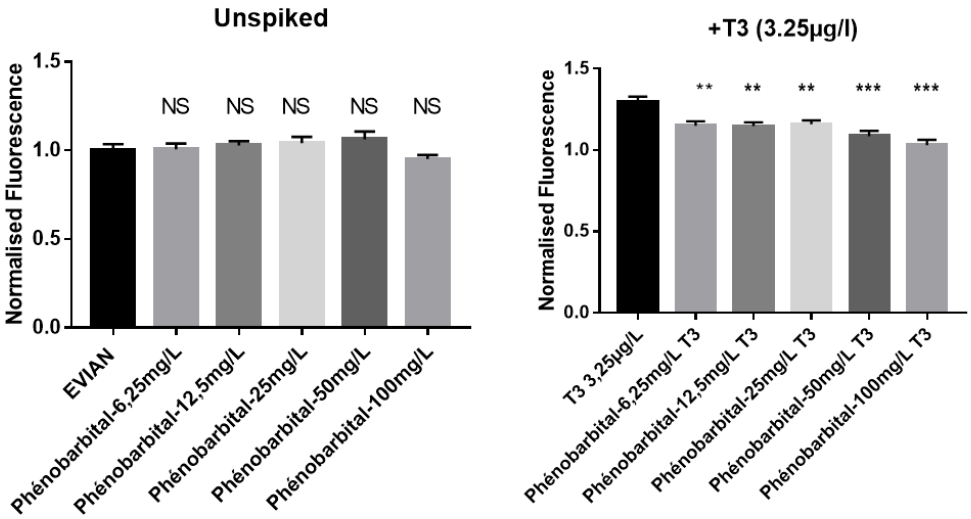

Figure 25 : XETA results for Phenobarbital.

4.6.3. Transgenic line availability

The THbZIP-GFP Xenopus is maintained in the lead lab (Watchfrog) and in different OECD countries in 3 public laboratories located in USA (University of Cincinnati), Portugal (Universidade de Aveiro) and France (Museum of Natural History) and one CRO located in Japan (IDEA consulting). Therefore 5 laboratories in total could provide and expand the transgenic line ensuring that the guideline could effectively be international and "long-lived". The THbZIP-GFP transgenic line is actually freely available from the laboratories describes below for collaborative research and commercially available for commercial uses.

- Laboratory of Dr Daniel Buchholz, McMicken College of Arts and Sciences, Department of Biological Sciences, University of Cincinnati, Cincinnati, USA:

- Laboratoire WatchFrog, Evry, France.
- Dr Yuta Onishi, Institute of Environnemental Ecology, IDEA Consultants, Inc., Japan.
- Dr Isabel Lopes, Departamento de Biologia & CESAM, Universidade de Aveiro, Portugal.
- Dr Jean-Baptiste Fini, Museum National d'Histoire Naturelle, Paris France..

4.6.4. Survival validation criteria

We decided to set the maximal allowed mortality in the control groups in the test guideline to 10% instead of 20% during the validation. This is supported by the results described in table 39 showing that this is effectively the limit expected for mortality as during the phase II of validation, 99% of the runs satisfied this criterion.

Survival		**80%**	**85%**	**90%**	**95%**	**100%**
Number of embryos (end of test)		16	17	18	19	20
Number of runs showing the indicated survival	Control	0	0	5	8	41
	T3 control	0	2	3	12	28
	T4 control	0	0	0	9	36
	Sum	0	2	8	29	105
Percentage of total runs (%)		**0**	**1**	**6**	**20**	**73**

Table 39: Survival in the control groups during the phase II of validation.

4.6.5. Maintenance of the eleutheroembryo in constant dark

During the WNT commenting round, a question was raised regarding the maintenance of the eleutheroembryo in constant dark during the test duration and the importance of phototaxis in frogs. Regarding phototaxis, anuran species are photopositive. This behavior allows the animal to stay in favorable feeding areas (Jaeger and Hailman, 1976), this behavior is therefore not important for the welfare of tadpoles in laboratory conditions, moreover working with developmental stages that don't need exogenous feeding. Rearing eleutheroembryo in the dark and under the XETA condition is not affecting their progression through developmental stage or survival, Elepfandt, (1996) also reported that blind individuals have been held without observed disadvantage in the same tank as sighted.

Elepfandt, A. (1996) 'Sensory perception and the lateral line system in the clawed frog' p97-120; in 'Biology of Xenopus' Eds Tinsley & Kobel: Oxford University Press,

Jaeger, R.G., and J.P. Hailman (1976) Ontogenic shift of spectral phototactic preferences in anuran tadpoles. J. Comp. Physiol. Psychol., 90:930-945.

4.6.6. Power analysis for a spacing factor of 10

In discussions during the OECD VMG-Eco meeting in October, 2018, VMG-Eco agreed to keep the maximum concentration of 100 mg/L and recommended to investigate if a spacing factor of 10 between concentrations could be used. With a larger spacing factor also low concentrations would be covered which can also be relevant for EDs effects, as well as high concentrations to make sure not to miss a potential effect on such a short exposure time. Therefore, it was requested to determine the impact of this modification on the power analysis for the XETA. A report containing the power analysis associated with this test design is included in Annex 4. The results show that it is also appropriate to use a factor of 10, the power to detect a 12% effect is maintained near 80% (76 to 81% depending on the simulated dose response shape). The power to detect a 15% effect exceeds 89% for all dose-response shapes considered.

5. REFERENCES

ASTM E1439-12, Standard Guide for Conducting the Frog Embryo Teratogenesis Assay-Xenopus (FETAX), ASTM International, West Conshohocken, PA, 2012, www.astm.org

Barney T Reed, May 2005. Guidance on the housing and care of the African clawed frog *Xenopus laevis*. Research Animals Department. *RSCPA*.

Castillo L, Seriki K, Mateos S, Loire N, Guédon N, Lemkine GF, Demeneix BA, Tindall AJ. In vivo endocrine disruption assessment of wastewater treatment plant effluents with small organisms. Water Sci Technol. 2013;68(1):261-8.

Demeneix B. Losing our minds : How chemical pollution affects the intelligence and mental health of future generations. Oxford series in behavioral neuroendocrinology, Oxford University Press, 2014.

Escher BI, Aït-Aïssa S, Behnisch PA, Brack W, Brion F, Brouwer A, Buchinger S, Crawford SE, Du Pasquier D, Hamers T, Hettwer K, Hilscherová K, Hollert H, Kase R, Kienle C, Tindall AJ, Tuerk J, van der Oost R, Vermeirssen E, Neale PA. Effect-based trigger values for in vitro and in vivo bioassays performed on surface water extracts supporting the environmental quality standards (EQS) of the European Water Framework Directive. Sci Total Environ. 2018 Jul 1;628-629:748-765

Fini, J. B., Le Mevel, S., Turque, N., Palmier, K., Zalko, D., Cravedi, J. P., and Demeneix, B. A. (2007). An in vivo multiwell-based fluorescent screen for monitoring vertebrate thyroid hormone disruption. Environ. Sci. Technol. 41, 5908–5914.

Fini, J. B., Dolo, L., Cravedi, J. P., Demeneix, B., and Zalko, D. (2009). Metabolism of the endocrine disruptor BPA by *Xenopus laevis* tadpoles. Ann. N. Y. Acad. Sci. 1163, 394–397.

Fini JB, Riu A, Debrauwer L, Hillenweck A, Le Mével S, Chevolleau S, Boulahtouf A, Palmier K, Balaguer P, Cravedi JP, Demeneix BA, Zalko D. Parallel biotransformation of tetrabromobisphenol A in *Xenopus laevis* and mammals: Xenopus as a model for endocrine perturbation studies. Toxicol Sci. 2012a Feb;125(2):359-67.

Fini JB, Le Mével S, Palmier K, Darras VM, Punzon I, Richardson SJ, Clerget-Froidevaux MS, Demeneix BA. Thyroid hormone signaling in the *Xenopus laevis* embryo is functional and susceptible to endocrine disruption. Endocrinology. 2012 Oct;153(10):5068-81.

Fini JB, Mughal BB, Le Mével S, Leemans M, Lettmann M, Spirhanzlova P, Affaticati P, Jenett A, Demeneix BA. Human amniotic fluid contaminants alter thyroid hormone

signalling and early brain development in Xenopus embryos. Sci Rep. 2017 Mar 7;7:43786.

Gray KM, Janssens PA. Gonadal hormones inhibit the induction of metamorphosis by thyroid hormones in *Xenopus laevis* tadpoles in vivo, but not in vitro. Gen Comp Endocrinol. 1990 Feb;77(2):202-11.

Green, Sherril L. (2009) *The Laboratory XENOPUS sp. CRC Press*.

Green JW, Springer TA, Holbech H 2018. Statistical Analysis of Ecotoxicity Studies. Wiley.

Kirillov A, Kistler B, Mostoslavsky R, Cedar H, Wirth T, Bergman Y. 1996. A role for nuclear NF-kappaB in B-cell-specific demethylation of the Igkappa locus. Nat Genet 13:435–441.

Kuiper GG, Klootwijk W, Morvan Dubois G, Destree O, Darras VM, Van der Geyten S, Demeneix B, Visser TJ. (2006). Characterization of recombinant *Xenopus laevis* type I iodothyronine deiodinase: substitution of a proline residue in the catalytic center by serine (Pro132Ser) restores sensitivity to 6-propyl-2-thiouracil. Endocrinology. 2006 Jul;147(7):3519-29

Leusch FDL, Aneck-Hahn NH, Cavanagh JE, Du Pasquier D, Hamers T, Hebert A, Neale PA, Scheurer M, Simmons SO, Schriks M. Comparison of in vitro and in vivo bioassays to measure thyroid hormone disrupting activity in water extracts. Chemosphere. 2018 Jan;191:868-875.

Leloup J & Buscaglia M 1977 Triiodothyronine, hormone of amphibian metamorphosis. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences Serie D 284 2261–2263.

Lim W, Nguyen NH, Yang HY, Scanlan TS, Furlow JD. A thyroid hormone antagonist that inhibits thyroid hormone action in vivo. J Biol Chem. 2002 Sep 20;277(38):35664-70..

Neale PA, Altenburger R, Aït-Aïssa S, Brion F, Busch W, de Aragão Umbuzeiro G, Denison MS, Du Pasquier D, Hilscherová K, Hollert H, Morales DA, Novák J, Schlichting R, Seiler TB, Serra H, Shao Y, Tindall AJ, Tollefsen KE, Williams TD, Escher BI. Development of a bioanalytical test battery for water quality monitoring: Fingerprinting identified micropollutants and their contribution to effects in surface water. Water Res. 2017 Oct 15;123:734-750.

O'Connor JC, Frame SR, Ladics GS. Evaluation of a 15-day screening assay using intact male rats for identifying antiandrogens. Toxicol Sci. 2002 Sep;69(1):92-108.

OECD 2012. Fish Toxicity Testing Framework, Series on Testing and Assessment no. 171. ENV/JM/MONO(2012)16. Chapter 3. http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=ENV/JM/MONO(2012)16&doclanguage=en

OECD 2006. Hypothesis Testing, in *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application,* Chapter 5. http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2006)18&doclanguage=en

USEPA 2003- Assessment of Pubertal Development and Thyroid Function in Juvenile Male CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 23 to 52/53 (http://www.epa.gov/scipoly/oscpendo/pubs/assayvalidation/ pubertal_male_ pr.htm)

USEPA 2004- Assessment of Pubertal Development and Thyroid Function in Juvenile Female CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 22 to 42/43 (http://www.epa.gov/endo/pubs/female_pubertal_report_feb_15_2004_rti.pdf)

Roth, P. (1941). Action antagoniste du propionate de testosterone dans la metamorphose experimentale des batraciens provoquee par thyroxine. Bull. Mus. Natl. Hist. Nat. (Paris) 13, 500-502.

Roth, P. (1942). Les antagonistes de la thyroxine dans la metamorphose des batraciens anoures la diiodotyrosine, le propionate de testosterone et le benzoate d'oestradiol. Bull. Mus. Natl. Hist. Nat. (Paris) 14, 48U83.

Roth, P. (1943). Action antagoniste du propionate de testosterone dans la metamorphose experimentale des batraciens anoures provoquee par thyroxine (2e note). Bull. Mus. Natl. Hist. Nat. (Paris) 15, 99-100.

Shi YB. Amphibian metamorphosis, From morphology to molecular biology, Whiley-liss 2000

Spirhanzlova P, De Groef B, Nicholson FE, Grommen SVH, Marras G, Sébillot A, Demeneix BA, Pallud-Mothré S, Lemkine GF, Tindall AJ, Du Pasquier D. Using short-term bioassays to evaluate the endocrine disrupting capacity of the pesticides linuron and fenoxycarb. Comp Biochem Physiol C Toxicol Pharmacol. 2017 Oct;200:52-58

Thompson, G. L. 1991. A note on the rank transform for interactions. Biometrika 78(3): 697–701.

Thompson G. L., Ammann L. P. 1989. Efficiencies of the rank-transform in two-way models with no interaction. Journal of the American Statistical Association. 4(405): 325–330.

Turque N, Palmier K, Le Mével S, Alliot C, Demeneix BA. (2005). A rapid, physiologic protocol for testing transcriptional effects of thyroid-disrupting agents in premetamorphic Xenopus tadpoles. Environ Health Perspect. 2005 Nov;113(11):1588-93.

Välitalo P, Massei R, Heiskanen I, Behnisch P, Brack W, Tindall AJ, Du Pasquier D, Küster E, Mikola A, Schulze T, Sillanpää M. Effect-based assessment of toxicity removal during wastewater treatment. Water Res. 2017 Dec 1;126:153-163.

Wang Z, Brown DD. Thyroid hormone-induced gene expression program for amphibian tail resorption. J Biol Chem. 1993 Aug 5;268(22):16270-8. .

Wegner S, Browne P, Dix D. Identifying reference chemicals for thyroid bioactivity screening. Reprod Toxicol. 2016 Oct;65:402-413.

Yen PM. Physiological and molecular basis of thyroid hormone action. Physiol. Rev. 2001 Jul;81(3):1097-142.

ANNEX 1: Modification of the incubation time

Modification of the incubation time

During the OECD validation phase, several experiments using T3 were undertaken to support a proposition to modify the incubation conditions from 72h at 21°C to 48h at 26°C. This modification presents several advantages, saving time and money when performing the XETA. The test duration is shortened, allowing to screen more test substances in a given time and cost. The cost of the chemical analysis will be decreased by one third.
These elements were discussed during the VMGeco meeting in 2018 and the experts panel decided that despites convincing results, a complete multilaboratory study would have been required to support this modification of the test design, given the potential additional stress and biological change that it could trigger.

First, a series of 15 T3 concentration-response experiments were performed at 72h (21°C) or 48h (26°C). The results (Figure 24) show the two concentration-response curves to be very close with an identical mean of fluorescence induction of 35% at 3.25 µg/L of T3.

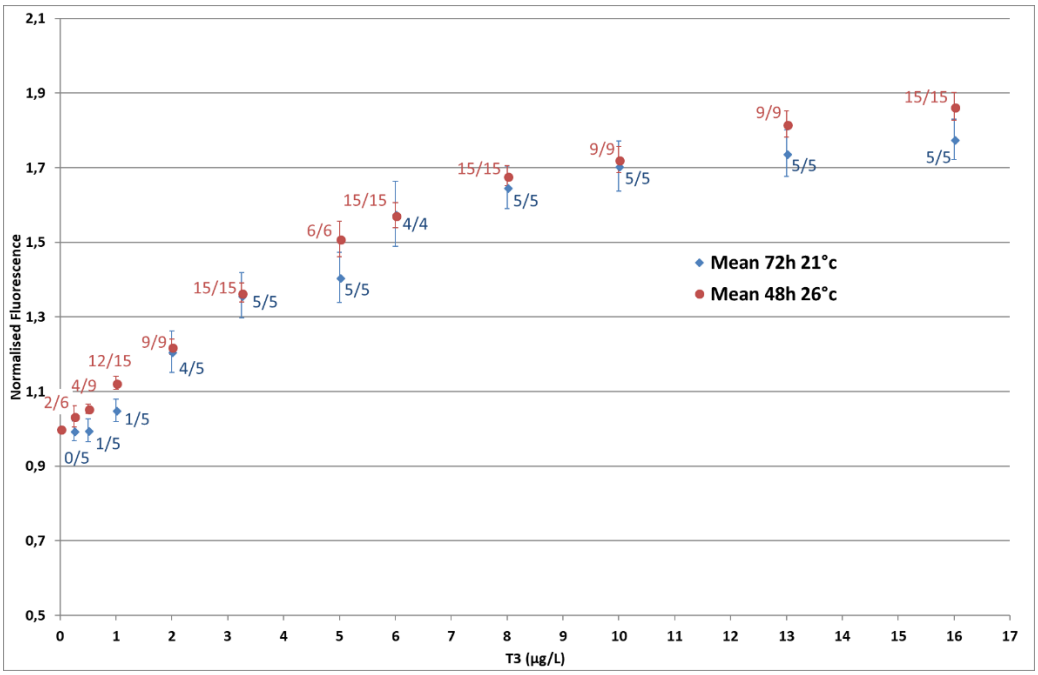

Figure 26 : Comparison of the concentration response of 12 concentrations of T3 after 72h incubation at 21°C or 48h incubation at 26°C.

The mean of the fluorescence and the SEM are represented. The fluorescence is normalized on the test medium control value. X/Y: Y indicate the total number of experiments performed for a given T3 concentration, X indicate the number of experiments showing a statistically significant induction of fluorescence compared to the test medium control.

Second, we performed 30 independent experiments over 6 months at 72h (21°C) or 48h (26°C) and compared the fluorescence inductions of the T3 control (Figure 26). Statistically identical results were obtained for the two conditions with a fluorescence induction of 41% at 48h (26°C) and 47% at 72h (21°C).

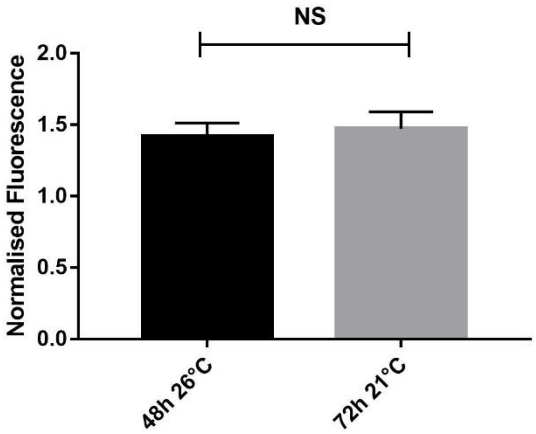

Figure 27 : Comparison of fluorescence induction in the T3 control after 72h incubation at 21°C or 48h incubation at 26°C.

The histogram represents the mean and SEM of 30 independent experiments.

ANNEX 2: Apparatus for fluorescence quantification

Spectrofluorimeter

Quantification of the fluorescence requires the use of a 96-well plate spectrofluorimeter equipped with bandpass fluorescence filters or a monochromator allowing the detection of GFP (max excitation wavelength 488 nm; maximal emission wavelength 510 nm). During the validation of the XETA the following 96-well plate spectrofluorimeter models were used successfully:

Model	TriStar LB941	Infinite200Pro	Synergy 2	Synergy 2
Manufacturer	Berthold	Tecan	Biotek	Biotek
Technology	Fluorescent filters	Fluorescent filters	Fluorescent filters	Monochromator

An example of SOP for the use of the TECAN infinite200Pro is provide below:

- Switch on the computer.
- Switch on the Tecan.
- Open the software ”i-control 1.10”.
- A window “Connect to instrument” opens. Click on “infinite200Pro”, then OK.
- Choose the following parameters:
  - Plate : [GRE96vt] Greiner 96 V transparent.
  -Part of plate: Choose wells if the plate is incomplete.
  - Fluorescence intensity:
    - Wavelength: 485 (20) nm
    - Mode: Top
    - Multiple Read/Well:
      Type: Square (filled)
      Size: 3*3
      Border: 0µM
    - Number of read :25
    - Gain: Manual→92
- Leave the other parameters as default.
- if the spectrofluorimeter has a temperature control module:
  - Set the temperature at 26.0 °C
  - Set the “Wait for temperature” parameters as Minimum-20.5°C and Maximum 21.5°C
- Put the plate in the machine. Press Start on the software.

- When the reading begins, an Excel file will open. Save it at the end of the first reading, and between each plate.

Fluorescent microscopy and imaging

Alternatively, fluorescence could be quantified using a fluorescence microscope equipped with filters and a camera suitable for fluorescence quantification. The following models have been used successfully:

Model	Olympus AX-70	Macrofluo
Manufacturer	Olympus	Leica
Technology	Bandpass GFP filters (excitation 475–495 nm, emission 495–545 nm)	Bandpass GFP filters (excitation 470/40nm,DM500, emission 525/50nm)
Objective	25x	Z6 Apo plan APO 1X
Additional features	None	Robotized platform (Märzhäuser)
Camera	Exi Aqa color camera (QImaging)	Black and white High-resolution ORCA-AG camera (Hamamatsu Photonics)
Imaging Software	QC Capture pro (QImaging)	SimplePCI (Hamamatsu Photonics)
Reference publications	1	2, 3, 4, 5

References

1: Fini JB, Mughal BB, Le Mével S, Leemans M, Lettmann M, Spirhanzlova P, Affaticati P, Jenett A, Demeneix BA. Human amniotic fluid contaminants alter thyroid hormone signaling and early brain development in Xenopus embryos. SciRep. 2017 Mar 7;7:43786.
2: Spirhanzlova P, De Groef B, Nicholson FE, Grommen SVH, Marras G, Sébillot A, Demeneix BA, Pallud-Mothré S, Lemkine GF, Tindall AJ, Du Pasquier D. Using short-term bioassays to evaluate the endocrine disrupting capacity of the pesticides linuron and fenoxycarb. Comp Biochem Physiol C Toxicol Pharmacol. 2017 Oct;200:52-58.
3: Neale PA, Altenburger R, Aït-Aïssa S, Brion F, Busch W, de Aragão Umbuzeiro G, Denison MS, Du Pasquier D, Hilscherová K, Hollert H, Morales DA, Novák J, Schlichting R, Seiler TB, Serra H, Shao Y, Tindall AJ, Tollefsen KE, Williams TD, Escher BI. Development of a bioanalytical test battery for water quality monitoring: Fingerprinting identified micropollutants and their contribution to effects in surface water. Water Res. 2017 Oct 15;123:734-750.
4: Välitalo P, Massei R, Heiskanen I, Behnisch P, Brack W, Tindall AJ, Du Pasquier D, Küster E, Mikola A, Schulze T, Sillanpää M. Effect-based assessment of toxicity removal during wastewater treatment. Water Res. 2017
5: Leusch FDL, Aneck-Hahn NH, Cavanagh JE, Du Pasquier D, Hamers T, Hebert A, Neale PA, Scheurer M, Simmons SO, Schriks M. Comparison of in vitro and in vivo bioassays to measure thyroid hormone disrupting activity in water extracts. Chemosphere. 2018 Jan;191:868-875.

ANNEX 3: October 2018 statistical report

October 10, 2018

Power Properties for XETA Studies,

John W. Green, Ph.D., Ph.D.
Principal Consultant Biostatistics, DuPont Data Science and Informatics

The current experimental design with 3 runs and 20 wells per run of each of 5 test concentrations and control. Several concentration-response shapes were simulated to capture the variety of trends observed in the validation studies. As indicated in a report sent in May, 2017, the XETA experimental design leads to variance components for run, run-by-treatment, and well nested within run-by-treatment. This is different from the ecotoxicity experimental designs used in most OECD guidelines where replicates refer to tanks, pens, or containers nested within each concentration. Analysis of XETA data treating runs incorrectly as nested within treatment has significant effects on the power properties of the tests. Only the Dunnett and Williams tests are readily analyzed with the correct variance structure. The power properties reported in this report contain results only for Williams' test and it is recommended that Williams' test be used when the data are consistent with a monotone concentration-response and Dunnett's test be used otherwise, following a transformation to achieve normality and variance homogeneity if necessary. Simulations were done that assumed test concentrations (treatments) were geometrically spaced with a constant ratio of 2. Phase 2 studies had somewhat different spacing (e.g., 1, 3, 10, 30, 100 ug/L) for some compounds, but the impact on power would be modest. Two sets of simulations were done.

In simulation set 1, the indicated variance components were obtained from Phase 2 using all 20 wells per treatment and run (i.e., no trimming). Simulations were done based on the range of variances so estimated and assuming an experimental design with 20 wells per treatment and run. These results are reported in Table 1.

In simulation set 2, the proposed protocol was followed whereby a 10% trim of the data from each run-treatment combination was done prior to analysis. This leads to omission of the 10% lowest FLUOR values and the 10% highest FLUOR values. In this set of simulations data from only 16 wells per run-treatment combination are used in the analysis. Variance components from the trimmed data were calculated and simulations were done based on the range of variances so estimated and assuming an experimental design with 16 wells per treatment (after trimming) and run. These results are also reported in Table 1.

By comparing the two power columns in Table 1, it will be clear that the trimming does not have a major impact on the power to find effects. It is well established that trimmed means are unbiased estimates of the true mean. References for that claim include

Johnson *et al.* (1994). Obviously, trimming will reduce the variance components, so the resulting analysis will have greater power to detect effects, that is, statistically significant differences between treatments and control. The reason for trimming the data prior to analysis has been explained in previous reports. As a reminder, sometimes an embryo is mistakenly placed in what is intended to be an empty well and sometimes an embryo is not placed in a well in some treatment group when there should be one there. Also, an embryo might die. It is difficult to verify these occurrences and they can skew the resulting estimated mean response in a given treatment or control. The number of such occurrences is expected to be small, so the 10% trimming is expected to eliminate them. Whether or not there are such problematic wells, trimming the data in this fashion does not bias the estimated mean response. Further discussion of this is given in Green et al. (2018).

It was observed in Phases 1 and 2, that some datasets failed the test for normality or variance homogeneity required by the Williams and Dunnett tests and no normalizing, variance stabilizing transform could be found. transform. In these cases, the NOEC obtained from the raw data or log- or square-root transformed data were almost always the same. A nonparametric analysis might seem preferable in these situations. However, there appears to be no way to do a nonparametric test that uses the correct variance-covariance structure and the power of these tests that ignore the variance-covariance issue tend to be very low. Some simulations were given that demonstrate this in the earlier report. It is preferable to use Williams' or Dunnett's test in these cases in spite of violation of the assumptions underlying those tests than to use a nonparametric test that ignores the experimental design. The error in the recommended protocol is much less than that in the nonparametric alternative and that can be demonstrated by comparing the power of the simulated parametric and nonparametric tests.

Not included in this report is the ability to estimate a specified percent effect (ECx) from a suitable regression model. That can be added later. However, the data from phase II indicates that successful regression modeling is only around 50% for the current design and observed variance.

Finally, the magnitude of variance observed in Phases I and II was simulated using the indicated variance structure. For Williams' test, a one-sided test was done since by its nature, a trend test, such as Williams is inherently one-sided. While most trends were observed to be increasing, some were decreasing. Simple examination of the data would dictate the direction of trend to test. This is analogous to regression model fitting, where a visual examination of the data would suggest fitting an increasing or decreasing concentration-response curve. It is possible to build into the regression model parameter estimation process a determination of the direction, but doing so has no effect on the sensitivity of the resulting model fit. It is possible to do a two-sided trend test, but doing so reduces the power by changing the p-values to 0.025 for each direction. See Green *et al*. (2018) or OECD (2006) for further discussion of two-sided trend tests. For Dunnett's test, two-sided tests were simulated since Dunnett's test is recommended when the data are not consistent with a monotone concentration-response, meaning changes in either direction should be considered. It should be emphasized that

consistency with monotone concentration-response does not mean the observed concentration-response must be monotone, only that it not deviate so strongly as to invalidate a test based on monotonicity. Williams' test is designed to smooth the deviations from monotonicity using the pool-the-adjacent-violators (PAVA) algorithm. The guidance provided earlier applies, namely that if three or more of the five treatment groups are amalgamated using PAVA, then concentration-response monotonicity is questionable and Dunnett's test should be used. Expert statistical judgment on a case-by-case basis may override this general guidance.

Four of the concentration-response shapes simulated are given in Figure 1. A fifth, designated ECPB=1, is not shown. It has a shape similar to that of ECPB=5 except the initial decline is even more rapid. The plots assume a maximum effect of 35%. The shapes will vary in predictable fashion according to the maximum effect simulated, which ranged from 30% to 5% for each shape. Power properties of the statistical tests are provided in Table 1. It will be observed that Williams' test has 90% power or greater to detect a 15% effect. By interpolation, 80% power should be achieved to detect a 12-13% effect.

Figure 1. Simulated concentration-response shapes

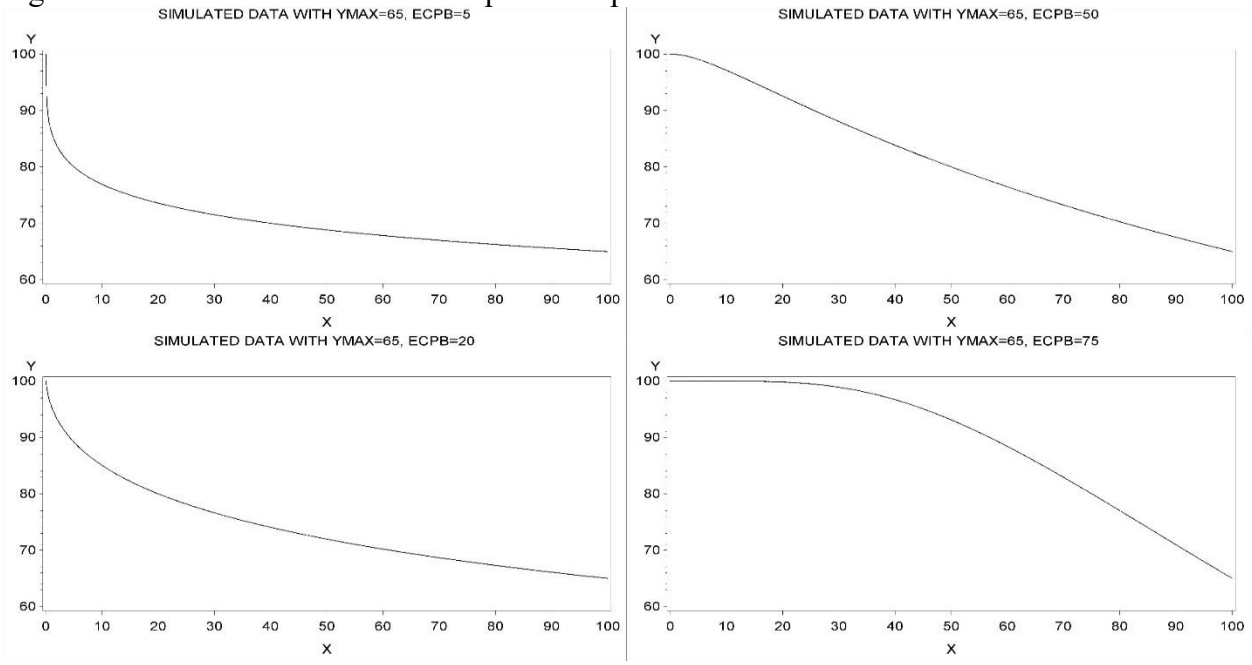

Figure 2. Power of tests for concentration-response shape with ECPB=5

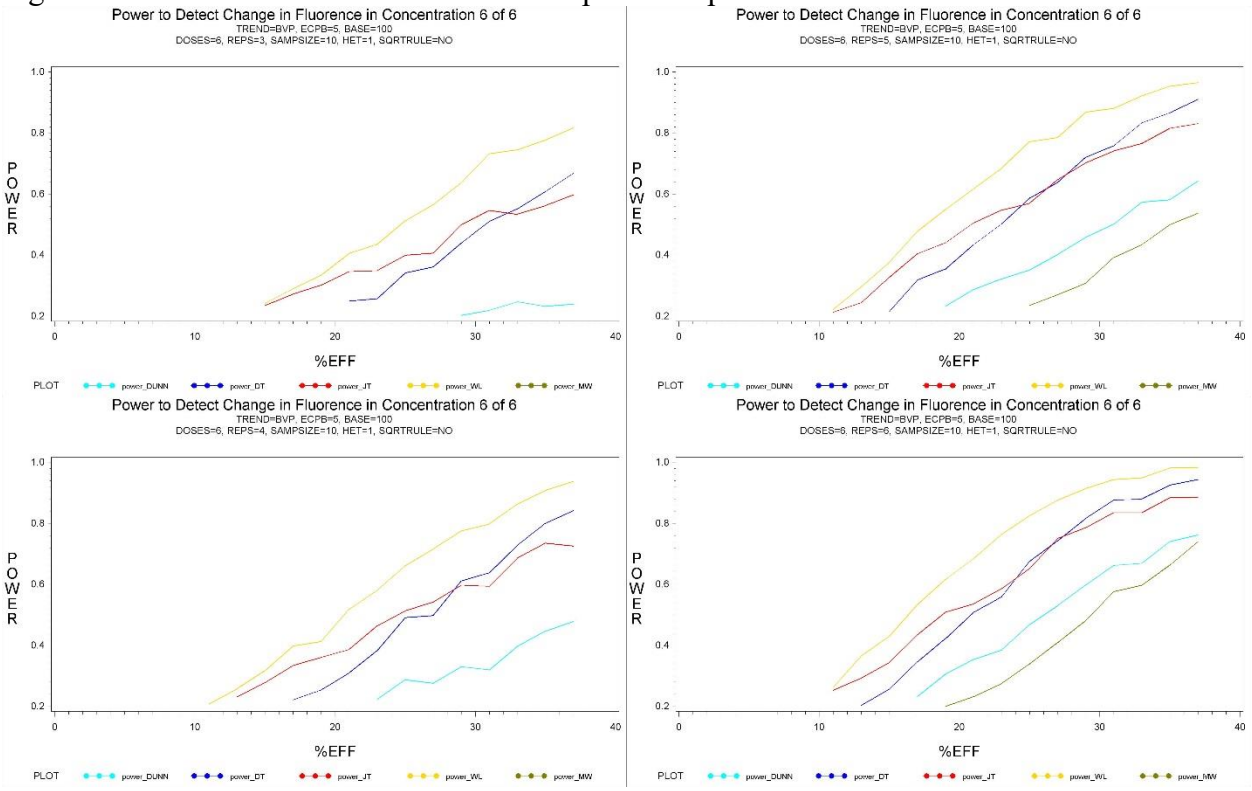

Table 1. Power of Williams and Dunnett tests for all concentration-response shape

		Power of Williams Test		Power of Dunnett Test	
ECPB	%Eff	Trimmed	Untrimmed	Trimmed	Untrimmed
1	12	78	73	68	72
5	12	78	70	76	82
25	12	77	69	74	72
50	12	78	69	83	72
75	12	80	69	88	83
1	15	91	82	97	95
5	15	90	81	96	96
25	15	91	83	97	97
50	15	90	81	92	96
75	15	92	81	75	74
1	20	99	94	99	86
5	20	98	93	98	95
25	20	98	94	98	96
50	20	99	96	89	96
75	20	99	94	95	97

The powers associated with trimmed data were based on 16 wells (after trimming) and the variance components observed from the trimmed data in Phase II.

The powers associated with untrimmed data were based on 20 wells (no trimming) and the variance components observed from the untrimmed data in Phase II.

It will be seen that when the 10% trimmed protocol is followed, Williams' test has 80% or greater power to detect a 12% effect for three simulated concentration-response shapes and 77-78% power to detect a 12% effect in the other two concentration-response shapes. The power properties of Williams' test with the trimmed analysis are consistent with the results of phase II. Of the 42 analyses in phase II, only 1 NOAEL from Williams test corresponded to more than a 12% effect. If no trimming is done, the power to detect a 12% effect falls below 70% for two concentration-response shapes, but is still higher than 80% for the other three simulated shapes. than concentration-response shapes simulated. There is general agreement within the statistical community that 80% power is adequate.

For the Dunnett test power results, it must be understood that the power is based on some treatment group being significantly different from the control. Depending on the shape of the concentration-response curve, as much as 10% of the simulated experiments had a low treatment significantly different from the control with the high treatment group not significantly different from the control. With greater variance, as in the untrimmed analysis, this was more likely. Williams' test is preferred since it captures the presumed monotone concentration-response and it is not possible for a low test concentration effect to be significant unless all higher concentration effects are also significant.

Further Comments on Nonparametric Analysis of XETA Studies

The obvious way to do a nonparametric analysis that takes the appropriate error structure into account would be to first do a rank-order transform, then perform Williams' or Dunnett's test or a general ANOVA on the transformed responses. However, there are several references that indicate that the errors rates, both positive and negative, for this type of analysis with complex structure tend to be very distorted. This is discussed in Green *et al*. (2018), where it is pointed out that Bathke and Lankowski (2005), Akritas *et al.* (1997), Akritas (1990) present simulation studies to show the rank-order transform can inflate the false positive rate for some multi-factor experimental designs. However Danbaba (2009, 2012) did a simulation study that indicated a normal score rank test did not suffer from this defect for some ANOVA type analyses. It seems prudent to avoid such rank-based analysis for the XETA experimental design until more conclusive evidence is available. To be clear, this is not an issue for a one-factor experiment analyzed at a single point in time, which is by far the most common experimental design in laboratory ecotoxicology studies. The rank transform is well validated and remains appropriate for those studies. But the XETA design is more complex.

Reducing the Number of Embryos Used in Testing

While it is desirable to minimize the number of embryos for the XETA test, the numbers recommended, 20 per treatment per run before trimming, is the minimum to obtain the necessary power across the possible concentration-response shapes that are likely to be

encountered. Simulation studies, not reported here, demonstrate that. In Table 1, it is already evident that the current sample size is barely adequate to detect a 12% effect.

References

Akritas MG, Arnold SF, Brunner E, 1997. Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs. Journal of the American Statistical Association 92: 258-265.

Akritas MJ 1990. The rank transform method in some two factor designs. JASA 85: 73-78.

Bathke A and Lankowski D 2005. Rank procedures for a large number of treatments. Journal of Statistical Planning and Inference 133: 223-238.

Danbaba A 2009. A Study of Robustness of Validity and Efficiency of Rank Tests in AMMI and Two-Way ANOVA Tests. Thesis University of Ilorin, Ilorin, Kwara State, Nigeria.
https://www.unilorin.edu.ng/graduatetheses/sc/Science_STA_2009_DANBABA.pdf

Danbaba, A. (2012). Comparison of a Class of Rank-Score Tests in Two-Factor Designs.
Nigerian Journal of Basic and Applied Science, 20 (4), pp 305-314

Green JW, Springer TA, Holbech H 2018. Statistical Analysis of Ecotoxicity Studies. Wiley. ISBN: 978-1-119-48881-1.

Johnson NL, Kotz S, and Balakrishnan N 1994. Continuous Univariate Distributions, volume 1, second edition. Wiley. ISBN:0-471-58495-9.

OECD 2006. *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application,*
http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2006)18&doclanguage=en

ANNEX 4: April 2019 statistical report

April 5, 2019

Power Properties for XETA Studies, extended

John W. Green, Ph.D., Ph.D.
Principal Consultant: John W Green Ecostatistical Consulting LLC

The current experimental design has 3 replicates and 20 wells per replicate of each of 5 test concentrations and control. Several dose-response shapes were simulated to capture the variety of trends observed in the validation studies. As indicated in reports sent in May, 2017, and October, 2018, the XETA experimental design leads to variance components for replicate, replicate-by-treatment, and well nested within replicate-by-treatment. This is different from the ecotoxicity experimental designs used in most OECD guidelines where replicates refer to tanks, pens, or containers nested within each dose. Analysis of XETA data treating replicates incorrectly as nested within treatment has significant effects on the power properties of the tests. Only the Dunnett and Williams tests are readily analyzed with the correct variance structure. The power properties reported in this report contain results only for Williams' test and it is recommended that Williams' test be used when the data are consistent with a monotone dose-response and Dunnett's test be used otherwise, following a transformation to achieve normality and variance homogeneity if necessary. Earlier simulations were done that assumed test concentrations (treatments) with approximate geometric spacing of 3.3. More accurately, the spacing was 1, 3, 10, 30, 100 ug/L. (In the October Power report, the simulations were at one point erroneously described as having a common ratio of 2.) In discussions of OECD VMG-eco in October, 2018, the question arose as to the impact on power of a design with geometric spacing with a common ratio of 10. This report contains the power properties associated with this alternative design. It will be observed that the power to detect a 12% effect is near 80% but does not consistently meet that objective. The power to detect a 15% effect exceeds 80% or all dose-response shapes considered.

In these simulation, the proposed protocol was followed whereby a 10% trim of the data from each replicate-treatment combination was done prior to analysis. This leads to omission of the 10% lowest FLUOR values and the 10% highest FLUOR values. In this set of simulations data from only 16 wells per replicate-treatment combination are used in the analysis. Variance components from the trimmed data were calculated and simulations were done based on the range of variances so estimated and assuming an experimental design with 16 wells per treatment (after trimming) and replicate. These results are reported in Table 1.

Not included in this report is the ability to estimate a specified percent effect (ECx) from a suitable regression model. That can be added later. However, the data from phase

2 indicates that successful regression modeling is only around 50% for the current design and observed variance.

Finally, the magnitude of variance observed in Phases 1 and 2 was simulated using the indicated variance structure. For Williams' test, a one-sided test was done since by its nature, a trend test, such as Williams is inherently one-sided. While most trends were observed to be increasing, some were decreasing. Simple examination of the data would dictate the direction of trend to test. This is analogous to regression model fitting, where a visual examination of the data would suggest fitting an increasing or decreasing dose-response curve. It is possible to build into the regression model parameter estimation process a determination of the direction but doing so has no effect on the sensitivity of the resulting model fit. It is possible to do a two-sided trend test, but doing so reduces the power by changing the p-values to 0.025 for each direction. See Green *et al*. (2018) or OECD (2006) for further discussion of two-sided trend tests. For Dunnett's test, two-sided tests were simulated since Dunnett's test is recommended when the data are not consistent with a monotone dose-response, meaning changes in either direction should be considered. It should be emphasized that consistency with monotone dose-response does not mean the observed dose-response must be monotone, only that it not deviate so strongly as to invalidate a test based on monotonicity. Williams' test is designed to smooth the deviations from monotonicity using the pool-the-adjacent-violators (PAVA) algorithm. The guidance provided earlier applies, namely that if three or more of the five treatment groups are amalgamated using PAVA, then dose-response monotonicity is questionable and Dunnett's test should be used. Expert statistical judgment on a case-by-case basis may override this general guidance.

Four of the dose-response shapes simulated are given in Figure 1. A fifth, designated ECPB=1, is not shown. It has a shape similar to that of ECPB=5 except the initial decline is even more rapid. The plots assume a maximum effect of 35%. The shapes will vary in predictable fashion according to the maximum effect simulated, which ranged from 30% to 5% for each shape. Power properties of the statistical tests are provided in Table 1. It will be observed that Williams' test has 90% power or greater to detect a 15% effect. By interpolation, 80% power should be achieved to detect a 12-13% effect.

Figure 1. Simulated Dose-response shapes

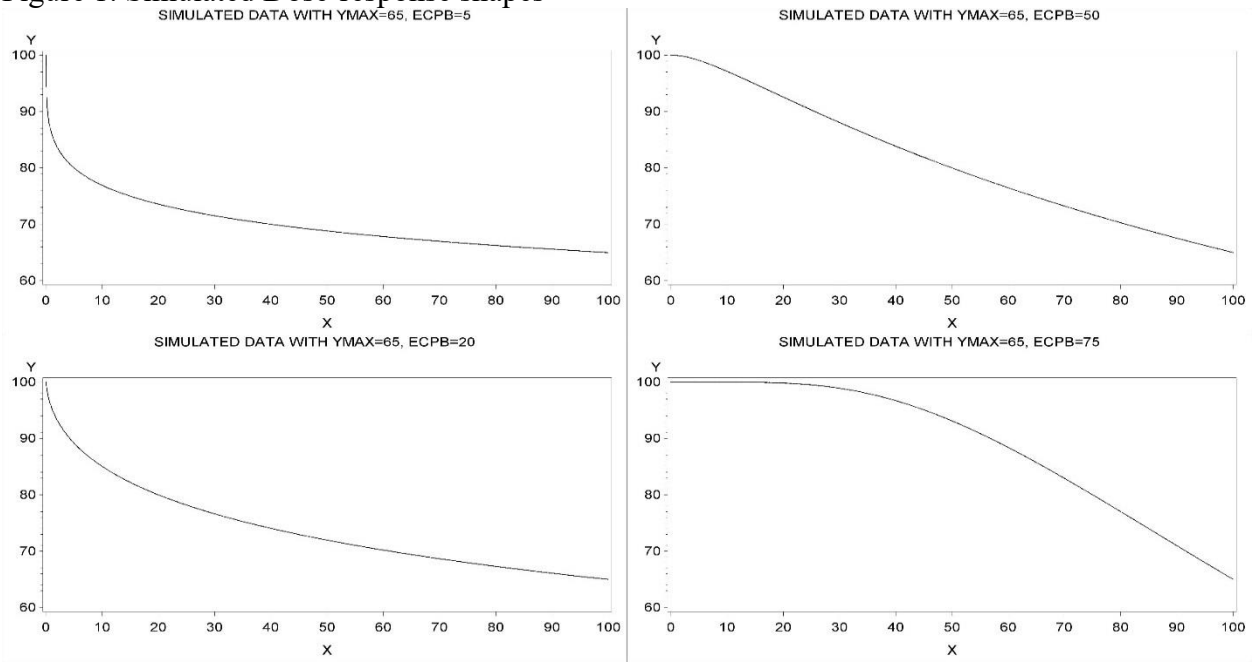

Figure 2. Power of tests for dose-response shape with ECPB=5

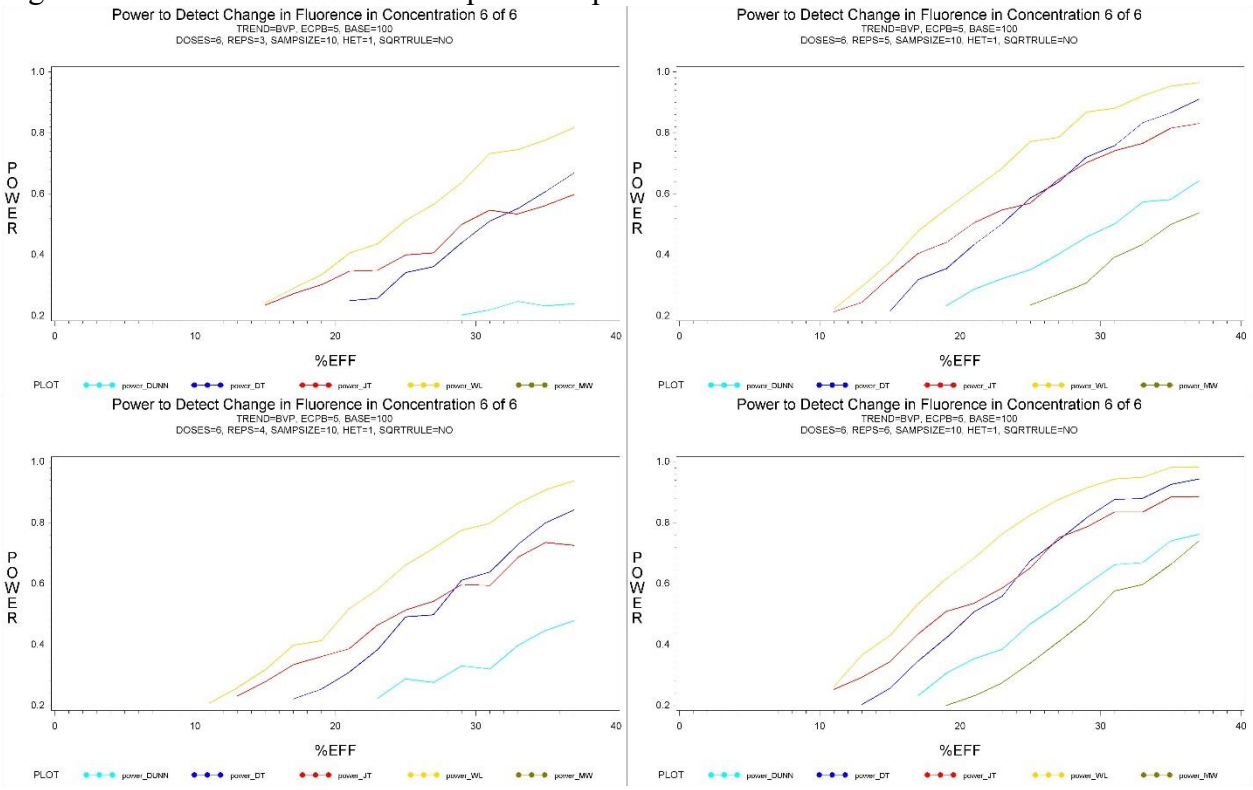

Table 1. Power of Williams test for all dose-response shape

ECPB	%Eff	WilPower
1	15	90
5	15	89
25	15	91
50	15	91
75	15	91
1	12	77
5	12	80
25	12	77
50	12	76
75	12	81

The powers based on 16 wells (after trimming) and the variance components observed from the trimmed data in Phase 2.

It will be seen that when the 10% trimmed protocol is followed, Williams' test has 89% or greater power to detect a 15% effect for all simulated dose-response shapes and 76-81% power to detect a 12% effect. The power properties of Williams' test with the trimmed analysis are consistent with the results of Phase 2. Of the 42 analyses in Phase 2, only 1 NOAEL from Williams test corresponded to more than a 12% effect.

References

Green JW, Springer TA, Holbech H 2018. Statistical Analysis of Ecotoxicity Studies. Wiley. ISBN: 978-1-119-48881-1.

OECD 2006. *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application,* http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2006)18&doclanguage=en