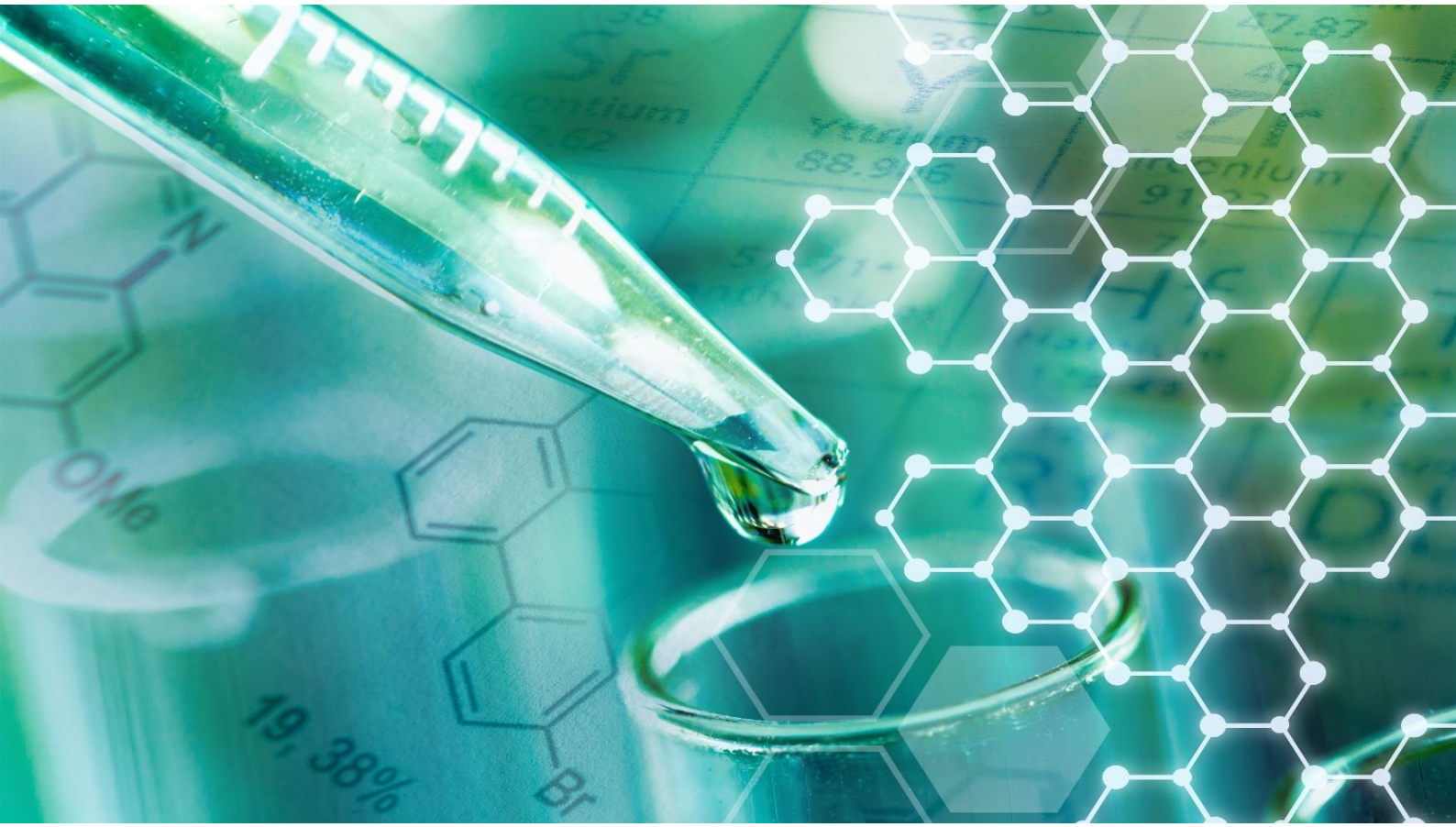
Thyroid in vitro methods: assessment reports by the thyroid disruption methods expert group

REPORTS ASSESSING THE VALIDATION STATUS OF ASSAYS FROM THE EU-NETVAL ACTIVITIES

Series on Testing and Assessment



OECD

Series on Testing and Assessment

Thyroid in vitro methods: assessment reports by the thyroid disruption methods expert group

Reports assessing the validation status of assays from the EU-NETVAL activities

Please cite this publication as:

OECD (2024), *Thyroid in vitro methods: assessment reports by the thyroid disruption methods expert group - Reports assessing the validation status of assays from the EU-NETVAL,* OECD Series on Testing and Assessment, OECD Publishing, Paris.

Photo credits: Cover © Billion Photos.

© OECD 2024

Attribution 4.0 International (CC BY 4.0)

This work is made available under the Creative Commons Attribution 4.0 International licence. By using this work, you accept to be bound by the terms of this licence (https://creativecommons.org/licenses/by/4.0/).

Attribution – you must cite the work.

Translations – you must cite the original work, identify changes to the original and add the following text: *In the event of any discrepancy between the original work and the translation, only the text of original work should be considered valid.*

Adaptations – you must cite the original work and add the following text: *This is an adaptation of an original work by the OECD. The opinions expressed and arguments employed in this adaptation should not be reported as representing the official views of the OECD or of its Member countries.*

Third-party material – the licence does not apply to third-party material in the work. If using such material, you are responsible for obtaining permission from the third party and for any claims of infringement.

You must not use the OECD logo, visual identity or cover image without express permission or suggest the OECD endorses your use of the work.

Any dispute arising under this licence shall be settled by arbitration in accordance with the Permanent Court of Arbitration (PCA) Arbitration Rules 2012. The seat of arbitration shall be Paris (France). The number of arbitrators shall be one.

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 38 countries in North and South America, Europe and the Asia and Pacific region, as well as the European Union, 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 Partner countries 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 twelve 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; Safety of Manufactured Nanomaterials;** and **Adverse Outcome Pathways.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (https://www.oecd.org/en/topics/chemical-safety-and-biosafety.html).

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, Basel, Rotterdam and Stockholm Conventions 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.

Foreword

For more than two decades, the OECD has been developing harmonised Test Guidelines, Guidance Documents, Detailed Review Papers for the screening and testing of endocrine disrupting chemicals. Several Test Guidelines are available for screening and testing chemicals for the estrogen, androgen, steroidogenesis mechanisms and their adverse effects on human health and the environment. A more limited choice of harmonised methods exists for the more complex thyroid hormone signalling pathway, especially no mechanistic assays have yet been developed into Test Guidelines. In 2014, the OECD published the New Scoping Document on *in vitro* and *ex vivo* assays for the identification of modulators of thyroid hormone signalling, followed by a call of the Working Party of the National Coordinators of the Test Guidelines Programme (WNT) to actually develop and validate the most advanced and relevant methods described in the Scoping Document. In 2017, EURL-ECVAM and the European Network of Validation Laboratories (EU-NETVAL) undertook a major endeavor to initiate the validation of seventeen *in vitro* assays. The work was mostly completed in 2022 and the Joint Research Centre published a report in 2023 (Bernasconi et al, 2023)[1] describing the different steps of the work and the outcome.

In April 2022, the WNT established an Expert Group on Thyroid Disruption Methods (TDM EG). The objectives of the group are to consolidate existing Test Guidelines, develop new ones, where appropriate, to determine mechanisms of action, and develop strategies for using various methods in combination in the most informative and efficient ways possible for the screening and testing of chemicals for thyroid disruption. One of the first tasks of the OECD Expert Group was to evaluate the validation status of the in vitro methods emanating from the EU-NETVAL initiative, and the level of readiness for standardization as a Test Guideline. Small size assessment groups were formed to go through documentation available for each assay, and address a common set of questions, including a blind assessment of the performance of the assay on the basis of data available from the EU-NETVAL laboratories.

Each assessment group studied rigorously the data generated, the SOPs and reports prepared by the EU-NETVAL laboratory and met at least three times via conference calls to reach consensus, if possible, on each question. The first assessment group met in October 2022. Five assays were completely assessed by May 2023; two more assays were assessed by October 2023; four more assays were assessed by April

[1] Full citation: Bernasconi C., Langezaal I., Bartnicka J., Asturiol D., Bowe G., Coecke S., Kienzler A., Liska R., Milcamps A., Munoz Pineiro A., Pistollato F., Whelan M., *Validation of a battery of mechanistic methods relevant for the detection of chemicals that can disrupt the thyroid hormone system*, Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/862948, JRC132532.

2024 and two more assays will enter the assessment pipeline in 2024. Assessment reports will be added as annexes to this compilation as they become available.

This document includes reports on the assessment of the validation status of individual thyroid in vitro assays from the EU-NETVAL activity in the period 2017-2022; assays are listed in chronological order of their progression in the assessment process.

Questions 16 and 17 concern conclusions reached by the assessment group (Q16) and recommendation for the next step(s) (Q17).

The assessment reports were presented to the entire TDM EG in May, October 2023 and June 2024. This report was presented to the WNT in April 2024 with the view to show the current state-of-play and encourage member countries to take action, i.e. offer leadership, and propose the development of Test Guidelines, and Guidance Documents, on single or combination of methods and organise the additional work recommended by the assessment group (e.g. transferability of the assay). The final report was submitted for WNT endorsement in July 2024.

Table of contents

Acknowledgements

The OECD secretariat would like to thank members of the TDM EG who volunteered to review data generated, reports prepared, respond to questions and meet and discuss to build a consensus:

Hakan Anderson, Lisa Baumann, Wieneke Bil, Olivier Blanck, Toine Bovee, Rhian Cope, Zhichao Dang, Sigmund Degitz, Chad Deisenroth, David du Pasquier, Jean-Baptiste Fini, Alexius Freyberger, Ellen Hessel, Miriam Jacobs, Mary Gilbert, Timo Hamers, Klára Hilscherová, Kostja Renko, Laurent Sachs, Rosemary Waring, Hilda Witters, Tom Zoeller.

Thanks are also addressed to EURL-ECVAM staff Ingrid Langezaal, Sharon Munn, Sandra Coecke, Roman Liska and Anne Milcamps who prepared material for the assessments, and were present in teleconferences of the assessment groups to respond to questions and collect feedback for the EU-NETVAL laboratories. Finally, thanks go to Mike Wade, Chair of the OECD Expert Group on Thyroid Disruption Methods, for his guidance throughout the process, and his knowledge of thyroid physiology shared throughout numerous conference calls.

1 OVERVIEW OF IN VITRO THYROID METHODS ASSESSED

Below is the list of assays emanating from the EU-NETVAL initiative assessed by volunteers from the OECD Expert Group. For a complete list of assays tested in the EU-NETVAL initiative, please refer to the JRC Technical Report (2023).

The study reports and SOPs on each of the assays are (or may become) readily available through the TSAR database maintained by EURL-ECVAM: https://tsar.jrc.ec.europa.eu/ .

Assay No. (Molecular Initiating event)	Assay Ttile (acronym) *Members of the assessment group*	NETVAL Laboratory
1b (TSH receptor)	Thyrotropin-stimulating hormone (TSH) receptor activation based on cAMP measurement *Chad Deisenroth, Hakan Andersson, Helena Hogberg-Durdock*	National Institute of Public Health - Centre of Toxicology and Health Safety (CZ)
2a (TPO inhibition)	Thyroid peroxidase (TPO) inhibition based on oxidation of Amplex UltraRed®. (TPO-AUR) *Sigmund Degitz, Alexius Freyberger, Mary Gilbert, Klára Hilscherova, Tom Zoeller*	RISE Research Institutes of Sweden (SE)
2c (Inhibition of tyrosine iodination)	Tyrosine iodination using liquid chromatography (TYRO-IOD) *Olivier Blanck, Sigmund Degitz, Alexius Freyberger, Klára Hilsherová*	Charles River Laboratories Den Bosch B.V. (NL)
3a (Binding to serum proteins TTR and TBG)	Thyroxine-binding prealbumin (TTR) / thyroxinebinding prealbumin (TBG) binding using fluorescence displacement (ANSA). (TTR-ANSA) *Olivier Blanck, Sigmund Degitz, Timo Hamers, Klára Hilscherová*	European Union Reference Laboratory for alternatives to animal testing / EURL ECVAM (IT)
3b (Binding to serum protein TTR)	Thyroxine-binding prealbumin binding using fluorescence displacement. (TTR FITC T4)	Wageningen Food Safety Research (NL)

Assay No. (Molecular Initiating event)	Assay Ttile (acronym) *Members of the assessment group*	NETVAL Laboratory
	Olivier Blanck, Sigmund Degitz, Timo Hamers, Klára Hilscherová	
4a (Inhibition of Deiodinase 1 activity)	Deiodinase 1 activity based on Sandell-Kolthoff reaction. (DIO 1)	BASF SE Experimental Toxicology and Ecology. Laboratory for Development of Alternative Methods (DE)
	Sigmund Degitz, Jean-Baptiste Fini, Miriam Jacobs, Kostja Renko	
4b (Inhibition of THs glucuronidation)	Inhibition of THs glucuronidation using liquid chromatography/mass spectrometry. (GLUC LC/MS).	Accelera S.r.l. (IT)
	Wieneke Bil, Olivier Blanck, Jean-Baptiste Fini, Rosemary Waring	
6a (TRα and TRβ receptor activation (agonist activity))	Human TH receptor alpha (TRα) and Human TH receptor beta (TRβ) reporter gene transactivation measuring agonist activities. (TRα and TRβ reporter assays)	Vitroscreen S. r. l. (IT)
	Hakan Anderson, Toine Bovee, Rhian Cope, Laurent Sachs, Hilda Witters	
6b (Human TRβ receptor (in)- activation (agonist and antagonist activity))	TR CALUX human TH receptor beta (TRβ) reporter gene transactivation measuring agonist and antagonist activities (TR CALUX)	Arpae / Vitrox (IT) (Method Developer BDS (NL) produced a second data set)
	Hakan Anderson, Toine Bovee, Rhian Cope, Laurent Sachs, Hilda Witters	
7a (Alteration of intrafollicular T4-content in Zebrafish eleutheroembryos)	Measurement of intrafollicular T4 using Zebrafish eleutheroembryos. (ZETA)	Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IT)
	Lisa Bauman, ZhiChao Dang, Chad Deisenroth, David du Pasquier. Ellen Hessel, Rosemary Waring	
8a (Alteration of cell proliferation of TH responsive cells)	T-screen assay measuring cell proliferation of GH3 cells using alamar blue/resazurin. (T SCREEN)	Nofer Institute of Occupational Medicine (PL)

Assay No. (Molecular Initiating event)	Assay Ttile (acronym) *Members of the assessment group*	NETVAL Laboratory
	Hakan Anderson, Toine Bovee, Rhian Cope, Laurent Sachs, Hilda Witters	

2 COMMON SET OF QUESTIONS POSED FOR EACH ASSAY

Each assessment group was asked the same set of questions and assessors were invited to respond to questions independently. The regular teleconferences (after Part 1, after Part 2 blind assessment, after part 2 unblinded assessment) allowed discussion and exchange among assessors and to reach consensus responses.

1- NON-blinded phase with reference and control chemicals (Part 1)

Part 1 report: Request for 5 valid runs for minimum the reference and control items of the method.

1. How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
2. Is the reproducibility across the runs consistent for the chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
3. How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
4. How do you qualify the dynamic range (signal to noise ratio) of the method? [*Consider the maximum response from the positive control or reference item and the minimum response from the solvent control*] [Excellent/Good/Fair/Poor]?
5. Any other observations on the method
6. Any other observations on the data

2- Blinded and unblinded phase (Part 2 report)

Part 2 report: Request for 3 valid runs for maximum 30 blind coded chemicals.

[Each assay had its own number of replicates, number of concentrations and number of chemicals tested in Part 2.]

2.1- When the chemicals are blinded:

7. How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]? [*This question can only be answered when more than 1 run is provided*]
8. Is the reproducibility across the runs consistent for the chemicals tested? In case you see an issue with one chemical, please flag it.
9. What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
10. For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]

2.2- When the chemicals are unblinded:

11. Now that the chemicals' identity is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak[2]/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
12. For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
13. How would you judge the specificity of the method?
14. Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
15. Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?

3- Conclusion and recommendation from the assessment group

16. What is the group conclusion on the validation status of the assay?
17. What further work (if considered necessary) the assessment group would recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method? or to confirm the acceptance criteria? Further development of the data interpretation procedure?)
18. For which chemicals is there sufficient information that they are active or inactive for the mode of action? Please indicate those, so that they can be considered for follow-up validation studies.

[2] The term "weak" was proposed in the questions distributed to the different assessment groups. However, over the course of the consensus building meetings, the definition of "weak" activity was discussed and no conclusion was reached. Preference was expressed to use terms "positive"/"active" (with or without qualifiers regarding strength of activity) or "negative"/"inactive", or "equivocal"

3 OVERVIEW OF CHEMICALS TESTED

The following table lists the chemicals tested in each *in vitro* assay. Most assays were tested using 30 chemicals selected based on available evidence of activity (or lack of activity) from human, animal or *in vitro* data from the literature.

The color code indicates the consensus conclusion from the assessment group on a particular assay based on the data assessed: **red** is the consensus for strongly active chemicals; **orange** is the consensus for weakly active chemicals; **yellow** is for equivocal chemicals, and **green** is for inactive chemicals, **grey** (not tested).

Although orange and red appear both in the table below, the assessment groups were not fully clear on the distinction to make between a “weakly active” versus a “clearly active” chemical, and all assessors indicated that a “weakly active” is in any case also a “clearly active” chemical (no ambiguity), but certain characteristics of the dose response might evoke activity at high concentrations only, for example.

In some cases, no clear consensus was reached and the diversity of assessment is represented by split of colours in the corresponding cell.

No.	Chemical name / CAS number	1b (TSHR)	2a (TPO-AUR)	2c (TYRO-IOD)	3a (TTR-ANSA)	3b; (TTR FITC-T4)	4a (DIO 1)	4b (GLUC-INH-LCMS)	6a (TRα TRβ assays)	6b (TR CALUX)		8a (T SCREEN)	7a. (ZETA)
										Ag.	Anta g.		
1	Mefenamic acid / 61-68-7												
2	PFOS / 1763-23-1												
3	2,4,6-Tribromophenol/ 118-79-6												
4	GC-1/Sobetirome / 211110-63-3												
5	6-Propyl-2-thiouracil / 51-52-5												
6	Silicristin / 33889-69-9												
7	Perchlorate (sodium)/ 7601-89-0												
8	BP2 / 2,2',4,4'-Tetrahydroxybenzophenone / 131-55-5												

No.	Chemical name / CAS number	1b (TSHR)	2a (TPO-AUR)	2c (TYRO-IOD)	3a (TTR-ANSA)	3b; (TTR FITC-T4)	4a (DIO 1)	4b (GLUC-INH-LCMS)	6a (TRα TRβ assays)	6b (TR CALUX)		8a (T SCREEN)	7a. (ZETA)
										Ag.	Anta g.		
9	3,3',5,5'-Tetrabromobisphenol A / 79-94-7												
10	Dibutyl phthalate / 84-74-2												
11	Aspirin / 50-78-2												
12	Pentachlorophenol / 87-86-5												
13	Triclosan / 3380-34-5												
14	Ampicillin / 69-53-4												
15	N,N,N',N'-Tetramethyl thiourea (TMTU) / 2782-91-4												
16	Ethylene thiourea / 96-45-7												

No.	Chemical name / CAS number	1b (TSHR)	2a (TPO-AUR)	2c (TYRO-IOD)	3a (TTR-ANSA)	3b; (TTR FITC-T4)	4a (DIO 1)	4b (GLUC-INH-LCMS)	6a (TRα TRβ assays)	6b (TR CALUX)		8a (T SCREEN)	7a. (ZETA)
										Ag.	Anta g.		
17	DiVanadium pentoxide / 1314-62-1												
18	Diclofenac / 15307-79-6												
19	Desipramin / 58-28-6												
20	Amiodarone / 19774-82-4												
21	Genistein / 446-72-0												
22	Salsalate / 552-94-3												
23	TETRAC / 67-30-1												
24	Ketoconazole / 65277-42-1												
25	Niflumic acid / 4394-00-7												
26	Sorafenib / 284461-73-0							T4G	T3G				

No.	Chemical name / CAS number	1b (TSHR)	2a (TPO-AUR)	2c (TYRO-IOD)	3a (TTR-ANSA)	3b; (TTR FITC-T4)	4a (DIO 1)	4b (GLUC-INH-LCMS)	6a (TRα TRβ assays)	6b (TR CALUX)		8a (T SCREEN)	7a. (ZETA)
										Ag.	Anta g.		
27	Cadmium chloride / 10108-64-2												
28	2-mercaptobenzothiazole / 149-30-4												
29	Resorcinol / 108-46-3												
30	Rosmarinic acid / 20283-92-5												
31	Morin hydrate xH2O / 654055-01-3												
32	Linolenic acid / 463-40-1												
33	Tannic acid / 1401-55-4												

No.	Chemical name / CAS number	1b (TSHR)	2a (TPO-AUR)	2c (TYRO-IOD)	3a (TTR-ANSA)	3b; (TTR FITC-T4)	4a (DIO 1)	4b (GLUC-INH-LCMS)	6a (TRα TRβ assays)	6b (TR CALUX)		8a (T SCREEN)	7a. (ZETA)
										Ag.	Antag.		
34	Fipronil sulphone / 120068-36-2												
35	Nordihydroguaiaretic acid / 500-38-9												
36	Hexadecyltrimethylammonium bromide / 57-09-0												
37	Linoleic acid / 60-33-3												
38	Bisphenol A diglycidyl ether / 1675-54-3												
39	Fipronil / 120068-37-3												
40	Octyl methoxycinnamate-2-Ethylhexyl 4-methoxycinnamate / 5466-77-3												

4 OVERVIEW OF CONCLUSIONS AND RECOMMENDATIONS FOR THE NEXT STEPS

Assay No. (Molecular Initiating event)	Conclusion and recommendation[3]
1b; TSH Receptor (Annex 1)	The current data do not allow to judge sufficiently the validation status of the assay, more positive chemicals are needed. Further identification of positive chemicals as reference chemicals in the agonism mode is needed. Development of the antagonism mode of the assay would be important, as well as identification of positive chemicals for the antagonism mode (e.g. a number of candidate drugs have been identified in PubMed as having a strong interaction with the TSH receptor). There are currently no chemicals identified for future validation activities with sufficient information.
2a; TPO-AUR (Annex 2)	The data provided show unacceptably high variability between runs and background responses, leading to large differences in results for the reference and control chemicals. The SOP for the AUR assay needs to be modified to improve the results. The EU-NETVAL laboratory ran the assay again with optimised SOPs using the following 5 chemicals: MMI, ETU, rosmarinic acid, triclosan, and 2,4,6-Tribromophenol. The data of the reference item improved, but there is still high variability of especially chemicals that seem to be inactive. In a second step, it would also be good to transfer the SOP to another lab. **Status 06/2024:** A final decision by the TDM-EG on whether or not the method is ready for transfer to another laboratory has yet to be taken.

[3] All chemicals CAS RN are available is section 5. Overview of Chemicals Tested.

Assay No. (Molecular Initiating event)	Conclusion and recommendation[3]
2c; TYRO-IOD (Annex 3)	Assay ready to be transferred to another laboratory; two more laboratories would provide more confidence; blind testing in another laboratory would also add confidence. Suitable chemicals were identified for further validation, with the understanding that not all of the chemicals need to be used, rather a sub-set.
3a; TTR-ANSA (Annex 4)	Further work is needed on the criteria for interference with the fluorescence readout. Once the SOP has been updated, transferability to at least one, better two, labs would be needed. A sub-set of Part 2 chemicals would need to be selected and blind tested (no sub-set of chemicals discussed within the assessment group).
3b; TTR- FITC T4 (Annex 5)	For the SOP: further work needed on the criteria for interference with the fluorescence readout. Generally the assessment group thinks the SOP for this assay is more ready than the SOP of the TTR-ANSA assay (3a). Further transferability to at least one, better two, labs would be needed. A sub-set of Part 2 chemicals would need to be selected and blind tested (no sub-set of chemicals discussed within the assessment group).
4a; DIO1 (Annex 6)	Assay ready to be transferred to at least one other laboratory; blind testing would also add confidence. Suitable chemicals were identified for further validation, with the understanding that not all of the chemicals need to be used, rather a sub-set of 2-5 from each category (clear inhibitor, weak inhibitor, negative).
4b; GLUC-INH-LCMS (Annex 7)	Assay seems to perform well (good reproducibility). The evaluation of specificity would benefit from testing additional negatives. Further transferability of the assay to at least one other laboratory would be needed to judge reproducibility. Some additional review of the UGT metabolites within the assay and the role of specific enzymes on its performance would be helpful in interpreting data generated in some cases/for some chemicals. Further testing of triclosan at the right dose levels would be important to confirm its utility in further transferability studies.
6a; TRα and TRβ receptor activation assays	The group concluded that the assay is valid for the detection of TRalpha and TRbeta agonists. There would be merit to investigate the utility of the assay for the detection of antagonistic activity in addition to agonistic activity.

Assay No. (Molecular Initiating event)	Conclusion and recommendation[3]
(Annex 8)	There are other ways (e.g. in silico) to identify chemical structures that are likely to bind the TR pocket. If there was a regulatory need/willingness to develop an OECD TG based on this method, between-lab variability would need to be evaluated. There was some discussion within the assessment group on the relative utility of this assay compared to *in silico* models. TETRAC and Sobetirome are the active chemicals and the rest are inactive.
6b; TR CALUX (Annex 9)	For agonism the assay looks specific. For antagonism the picture is less clear and further experimental data would be needed to conclude on the assay performance. There are other ways (e.g. *in silico*) to identify chemical structures that are likely to bind the TR pocket. If there was a regulatory need/willingness to develop an OECD TG based on this method, between-lab variability would need to be evaluated. There was some discussion within the assessment group on the relative utility of the agonist part of this assay compared to *in silico* models. For the antagonistic part of the assay, there is not enough knowledge to be able to say anything about replacement by in silico models. For the agonistic part of the assay: TETRAC and Sobetirome are the active chemicals and the rest are inactive. For the antagonistic part of the assay: mefenamic acid was the only chemical considered to be positive.
8a; T-SCREEN (Annex 10)	The group concluded that the assay is valid for the detection of TR agonists, but the response measured (cell proliferation) is unspecific of thyroid modalities. Other hormones and growth stimulators would need to be tested to assess the specificity of the assay. TETRAC and sobetirome are clear positive chemicals.
7a; ZETA (Annex 11)	The conclusion of the assessment group is that the assay is promising but more work is needed. - Number of embryos should be increased; - More negative chemicals should be tested to ascertain the specificity of the assay; - The concentrations tested and solvent concentration should be harmonised with other TG 236-like assays (based on LC10 for example for the MTC)

Assay No. (Molecular Initiating event)	Conclusion and recommendation[3]
	- The spacing of concentrations should be reevaluated and probably be spaced by 1/2Log, or best based on a range finding test to determine toxicity. - Data interpretation procedure for the statistical treatment of data from the 3 runs need further thinking; - Depending on the problem formulation (context of use), this assay might be expanded (testing a fourth concentration,..) for a better dose-response characterisation. - More work is needed on the source of antibodies. It is difficult to recommend any negative chemicals at this stage based on the results obtained. For the positive chemicals, the top four chemicals recommended are resorcinol, perchlorate, PTU and mefenamic acid, but all chemicals tested would do a good job as positive chemicals.
8a;T SCREEN (Annex 10)	The group concluded that the assay is valid for the detection of TR agonists, but the response measured (cell proliferation) is unspecific of thyroid modalities. Other hormones and growth stimulators would need to be tested to assess the specificity of the assay. TETRAC and sobetirome are clear positive chemicals.

In summary, for the promising methods among the assays assessed above, the TDM-EG recommends to proceed with validation as follows:

- Transfer the method to 1 or 2 other laboratories to demonstrate the method can be performed by others.
- The laboratory first performs a study with the reference and control chemicals to show acceptance criteria in the SOP can be met.
- The laboratory then performs a study with around 10 blind-coded chemicals having a good balance between positive/negative/equivocal expected outcomes.

5 ANNEXES

Annex 1- Thyrotropin-stimulating hormone (TSH) receptor activation based on cAMP measurement (TSHR)

Annex 2- Thyroid peroxidase (TPO) inhibition based on oxidation of Amplex UltraRed®. (TPO-AUR)

Annex 3- Tyrosine iodination using liquid chromatography (TYRO-IOD)

Annex 4- Thyroxine-binding prealbumin (TTR) / thyroxine-binding prealbumin (TBG) binding using fluorescence displacement (ANSA). (TTR-ANSA)

Annex 5- Thyroxine-binding prealbumin binding using fluorescence displacement. (TTR FITC T4)

Annex 6- Deiodinase 1 activity based on Sandell-Kolthoff reaction.

Annex 7- Inhibition of thyroid hormones (THs) glucuronidation using liquid chromatography/mass spectrometry (LC/MS)

Annex 8- Human thyroid hormone receptor alpha (TRα) and Human thyroid hormone receptor beta (TRβ) reporter gene transactivation measuring agonist activities

Annex 9- CALUX human thyroid hormone receptor beta (TRβ) reporter gene transactivation measuring agonist and antagonist activities

Annex 10- Measurement of proliferation of rat pituitary-derived cell line GH3

Annex 11- Measurement of intrafollicular thyroxine (T4) using zebrafish eleutheroembryos

Glossary of abbreviations used:

BP2: 2,2',4,4'-Tetrahydroxybenzophenone (CAS RN. 131-55-53)

ETU: Ethylene thiourea (CAS RN.96-45-7)

MMI: 1-méthyl-3H-imidazole-2-thione (Methimazole) (CAS RN. 60-56-0)

PTU: 6-propyl-2-thioxo-2,3-dihydropyrimidin-4(1H)-one (Propylthiouracil) (CAS RN. 51-52-5)

QLI: Quanti-Lum luciferase inhibition assay

TPO: Thyroperoxydase

ANNEX 1- Thyrotropin-stimulating hormone (TSH) receptor activation based on cAMP measurement (TSHR)

NON-blinded phase with reference and control chemicals (Part 1)

TSHR Part 1 report: 4 runs are available for DMSO as a test item and TSH as a reference item with 7 concentrations/chemical.

Question 1	How do you qualify reproducibility across the 4 runs? [Excellent/Good/Fair/Poor]?
Consensus	Preliminary consideration: the EU-NETVAL laboratory developed the SOPs from scratch, without support from a method developer. Generally the assessment group was in agreement that there are areas for improvement and made some concrete suggestions, should the assay be taken forward. A source of inspiration to reduce variability might be the corresponding ToxCast/Tox21 assays (shorter incubation times, intracellular cAMP measured) In response to the question 1: only the TSH response can truly be evaluated in terms of variability across runs. DMSO is a solvent and negative control, rather than a true test item with anticipated bioactivity. The intra-experimental variability, based on OD values between replicates appeared fairly low. The inter-experimental variability (between runs) was high, however, suggesting poor reproducibility.
Question 2	Is the reproducibility across the 4 runs consistent for the 2 chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	In the case of this assay, questions 1 and 2 produce the same answers.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	Generally, the replicates within each run looked consistent, based on the raw plate reader OD values. Intra-experimental variability is considered to be good.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	Within each run, the dynamic range is good, but when comparing between runs, there is a large difference in the dynamic range of the assay.
Question 5	Any other observations on the method
Consensus	The following points were consensual: 1. Variability of the method may be lowered by reducing the exposure duration to approximately 0.5-2 hours and measuring intracellular [cAMP] rather than secreted [cAMP];

	2. DMSO is a solvent, not a control or test item; another chemical should be chosen as negative control if possible, another hormone such as T4 was suggested; 3. Amend the calculation of results by interpolating the % bound of each well to the cAMP concentration. 4. Set a defined concentration series specifying concentration intervals and range, would standardize the data interpretation. 5. Develop the antagonist version of the assay to expand relevance to the expected more prevalent chemical mode-of-action. 6. Since the assay is coupled to a viability endpoint (NRU assay), the cytotoxicity data could be incorporated into a DIP, eliminating the need for an initial range-finding assay (MTT assay).
Question 6	Any other observations on the data
Consensus	One suggestion to possibly reduce variability was to carry out a normalisation of data to the positive control.

Conclusion from the assessment group on Part 1 report?
[Nothing specific was mentioned by the assessment group]

Recommendations from the assessment group for further standardisation of method?
[Nothing specific was mentioned by the assessment group]

Part 2 report when the chemicals are blinded:

Part 2 report: 1 run are available for 30 chemicals tested, with 7 concentrations/chemical.

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Not applicable. The reason for having only one run is cost-efficiency, it was decided to have one run only for the negative results and three runs for positive results.
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	Not applicable.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	Much more data would be needed to determine the cut-off. Some default could be to apply a 3x standard deviation, or a 20% activation, but the group did not feel confident providing any definitive answer.
Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	All chemicals resulted to be negative in the assay. There was some discussion about whether concentrations tested were high enough, and it seems that yes the concentrations were comparable to other assays in terms of molarity (to be checked), thus relevant and avoiding cytotoxicity.

Part 2 report when the chemicals are unblinded:

Question 11	Now that chemicals identity is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	<table><tr><th>Chemical</th><th>CD</th><th>HH</th><th>HA</th></tr><tr><td>Mefenamic acid</td><td></td><td></td><td></td></tr><tr><td>PFOS</td><td></td><td></td><td></td></tr><tr><td>2,4,6-Tribromophenol</td><td></td><td></td><td></td></tr><tr><td>GC-1/Sobetirome</td><td></td><td></td><td></td></tr><tr><td>6-Propyl-2-thiouracil</td><td></td><td></td><td></td></tr><tr><td>Silicristin</td><td></td><td></td><td></td></tr><tr><td>Perchlorate (sodium)</td><td></td><td></td><td></td></tr><tr><td>2,2',4,4'-Tetrahydroxybenzophenone</td><td></td><td></td><td></td></tr><tr><td>3,3',5,5'-Tetrabromobisphenol A</td><td></td><td></td><td></td></tr><tr><td>Dibutyl phthalate</td><td></td><td></td><td></td></tr><tr><td>Aspirine</td><td></td><td></td><td></td></tr><tr><td>Pentachlorophenol</td><td></td><td></td><td></td></tr><tr><td>Triclosan</td><td></td><td></td><td></td></tr><tr><td>Ampicilin</td><td></td><td></td><td></td></tr><tr><td>N,N,N',N'-Tetramethyl thiourea (TMTU)</td><td></td><td></td><td></td></tr><tr><td>Ethylene thiourea</td><td></td><td></td><td></td></tr><tr><td>DiVanadium pentoxide</td><td></td><td></td><td></td></tr><tr><td>Diclofenac</td><td></td><td></td><td></td></tr><tr><td>Desipramin</td><td></td><td></td><td></td></tr><tr><td>Amiodarone</td><td></td><td></td><td></td></tr><tr><td>Genistein</td><td></td><td></td><td></td></tr><tr><td>Salsalate</td><td></td><td></td><td></td></tr><tr><td>TETRAC</td><td></td><td></td><td></td></tr><tr><td>Ketoconazole</td><td></td><td></td><td></td></tr><tr><td>Niflumic acid</td><td></td><td></td><td></td></tr><tr><td>Sorafenib</td><td></td><td></td><td></td></tr><tr><td>Cd chloride</td><td></td><td></td><td></td></tr><tr><td>2-mercaptobenzothiazole</td><td></td><td></td><td></td></tr><tr><td>Resorcinol</td><td></td><td></td><td></td></tr><tr><td>Rosmarinic acid</td><td></td><td></td><td></td></tr></table> The test design of this assay did not allow to evaluate the performance of the method using the chemicals tested (all chemicals tested negative). The assessment group thought that possibly ETU could have turned to be positive. ETU data for the agonism mode from the Tox21 database was shared and although not very convincing, there was a trend. Amiodarone was flagged as possible positive for the agonism mode during the chemical selection process, but no support in literature.
Question 12	For each of the chemicals that were flagged as ‘inconsistent between runs’, can you identify/hypothesise the reason?
Consensus	Not applicable.
Question 13	How would you judge the specificity of the method?
Consensus	Not applicable.

Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	Not possible to make any good recommendation at this stage given the current set of data.
Question 15	Are 2 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	Not possible to make any good recommendation here given all single runs are negative for all chemicals. In order to evaluate specificity/sensitivity, it is necessary to have more runs (especially for positive chemicals, but also to confirm negatives) to be able to judge how many runs might be needed in routine testing.

Conclusion and recommendation from the assessment group:

Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The current data do not allow to judge sufficiently the validation status of the assay, more positive chemicals are needed.
Question 17	What further work (if considered necessary) the assessment group would recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method? or to confirm the acceptance criteria? Further development of the data interpretation procedure?)
Consensus	Further identification of positive chemicals as reference chemicals in the agonism mode. Development of the antagonism mode of the assay would be important, as well as identification of positive chemicals for the antagonism mode (e.g. a number of candidate drugs have been identified in PubMed as having a strong interaction with the TSH receptor).
Question 18	For which chemicals there is sufficient information that they are active or inactive for the mode of action? Please indicate those, so that they can be considered for follow-up validation studies.
Consensus	There are no chemicals currently with sufficient information.

ANNEX 2- THYROID PEROXIDASE (TPO) INHIBITION BASED ON OXIDATION OF AMPLEX ULTRARED®. (TPO-AUR)

NON-blinded phase with reference and control chemicals (Part 1)

TPO-AUR Part 1 report: 5 runs are available for a few chemicals tested including the reference and control items of the method and 10 additional chemicals.

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	Poor, and very variable across chemicals. The results of the reference item MMI are not well reproducible either. The assessment group is only assessing the activity of TPO here.
Question 2	Is the reproducibility across the 5 runs consistent for the 7 chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	The assessment group was puzzled as there were major differences in between-run reproducibility between the chemicals, especially the performance using the reference chemical was not convincing (large spread of data). The group was concerned about whether the poor performance/very variable baseline is related to poor execution by the lab (accumulation of technical errors?) or to the source of the TPO enzyme (human versus rat in the test developer's lab) or to some steps in the preparation/extraction of the enzymes, or to the normalisation of the baseline data? The SOPs should prescribe the normalisation (each batch has to pass a validation of protein content before adding to the wells). The group questioned the need for evaluating cytotoxicity. Note: The lab indicated that they used 7 different batches of cell lysates without pooling them, which could be a major source of variability. The MMI was not calibrated across the different experiments. BP2 had consistent results across the runs, and not the other chemicals, which is puzzling the assessment group.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	Variability within each run was very good.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	Good, but variable. There were different interpretations of what "dynamic range" is (sensitivity being a component of the dynamic range).
Question 5	Any other observations on the method
Consensus	The experts questioned the relevance of both the cytotoxicity and the luciferase assays because it does not make much sense to evaluate the non-specific enzyme inhibition in this assay nor the cytotoxicity in a cell-free assay.

	Large variability across chemicals is a concern for the assessment group. Need to investigate with the lab what happened (several hypotheses made by the assessment group).
Question 6	Any other observations on the data
Consensus	The group would be interested to know which batch of cell lysate was used for which chemicals and whether that could explain the variability across chemicals. This information is not available in the Part 1 report. Comments were made about the acceptance criteria (3 orders of magnitude) too large, usually not more than one order of magnitude or historical mean +/- 2SD.

Conclusion from the assessment group on Part 1 report?
Recommendations from the assessment group for further standardisation of method?

The assessment group noted they would like to see more than just summary statistics (especially CVs and SDs).

It seems that invalid runs have nevertheless been used (?) and further clarity would be needed.

More clarity needed on the batches (pooled? Not pooled?) Performance of each batch?
Raw data on the control with and the control without enzyme (blank) performance in relation to each batch would be useful to see.

After the meeting, the following was confirmed by the EU-NETVAL laboratory:

- Different batches of cell lysates with TPO were used during the studies. They were not pooled.
- CVs and SDs for triplicate samples were added to the data analysis, and information on which results were created with which batch were provided.
- With one exception, no invalid runs were used for the final results. Information on valid and invalid runs is provided with the report.

Additional information provided by the test developer at USEPA

1. Regarding the cytotoxicity assay, this assay was included to understand the chemical library as plated and was a part of our broad screening exercise. This gave us some insight into the potential reactivity/cytotoxicity of chemicals as plated, thereby giving some context for concentrations observed to be positive in the AUR-TPO assay. Understanding concentrations that might be cytotoxic could help prioritize chemicals with highly specific and potent activity toward TPO when screening. As you and the panel note it is not strictly necessary for conduct of the AUR-TPO assay but as part of a screening exercise it was something useful to us. Perhaps it could be deleted from your SOP or included as an optional appendix.
2. Regarding the QLI assay, this was another assay run for our screening exercise to provide context for the observations in the AUR-TPO assay, i.e., to identify chemicals that may be nonspecific protein inhibitors. It seemed unlikely to us that a chemical would both specifically inhibit QLI and TPO, though possible, and as such concentrations that appear active in both the AUR-TPO assay and the QLI assay may suggest a chemical that acts as a detergent, salt, reactive, etc. Thus, the QLI assay is not indicative of interference with TPO. Its inclusion is to try to understand the potential for nonspecific hits in the AUR-TPO assay since we are using an indirect measure (fluorescent light signal from AUR product) to indicate potential TPO inhibition.

Part 2 report when the chemicals are blinded:

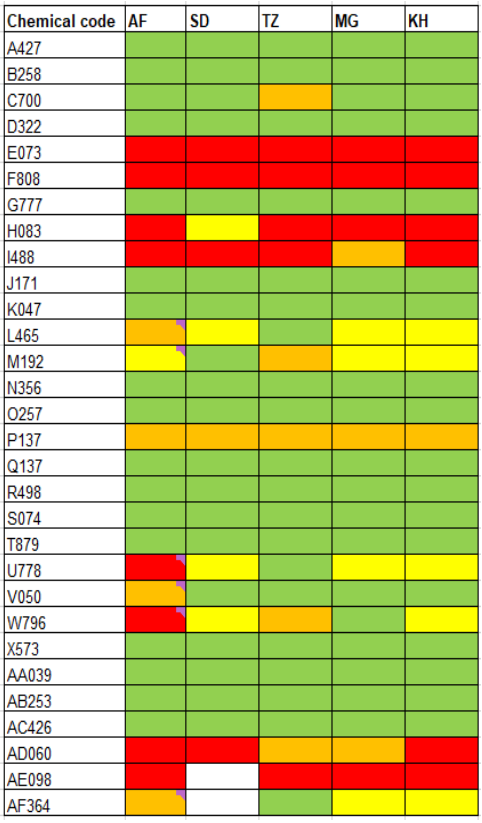
Part 2 report: 3 runs are available for 30 chemicals tested, with 8 concentrations/chemical.

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	There was agreement that when the tested chemical is active (clear inhibitor), there is good reproducibility across the runs or experiments. When the tested chemical is inactive or only active at low concentrations, there can be strong variability across the runs or experiments and reproducibility is poor. A phenomenon of stimulatory effects above control values was observed for certain tested chemicals, which could potentially be masking inhibitory effects at higher concentrations. An hypothesis was that these chemicals could be surfactants that interact with the preparation to enhance the activity of the enzyme and create background noise. It was observed that the start of the dose-response curves at the lowest concentrations tested varied between -40% to +20% for the test items. Also the % inhibition by the lowest concentration of the reference item MMI varied substantially between -35% to +10%. The group discussed possible causes for this variability: • Is there a batch related effect due to the preparation of the enzyme or should the 9 months time between the preparation and experiment be reduced? Were the cells scraped or trypsinised? Trypsin is known to inhibit TPO. • Is there a difference in values that is related to the position of samples on the plate? This was questioned because the solvent control sample on the right of the plate was often lower than the lowest MMI concentration. Is the plate reader a single well reader? Further investigation into the cause of variability is necessary and findings should be captured in an improved SOP. A criterion could be included for the start of the curve e.g. between -10% and +10%.
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	Similar to the previous question, there was agreement that there is good reproducibility across the runs in cases the tested chemicals were clear inhibitors of TPO (3 chemicals), and strong variability in case of inactive chemicals. The reference substance MMI also showed strong variability across the runs.

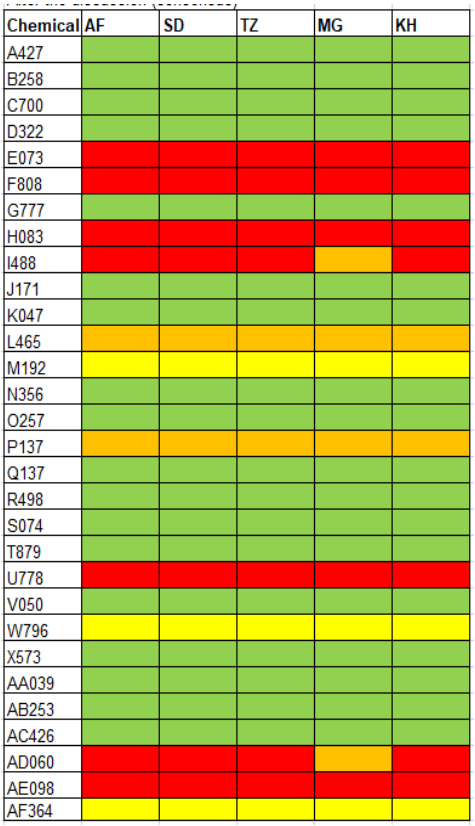
<table>
<tr><td>Question 9</td><td colspan="2">What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)</td></tr>
<tr><td>Consensus</td><td colspan="2">There was agreement that the cut-off value to identify inhibitors of TPO is 20%inhibition.

There was concern however in case where the baseline activity is negative, as was observed in some cases where the low concentrations seemed to indicate stimulation of TPO rather than inhibition. This could result in false-negative outcome of the assay. One assessor was of the opinion that the negative baseline resulted from an execution error in the protocol.

There was agreement that further investigation is needed with the laboratory to resolve the variability of the lowest concentrations tested.</td></tr>
<tr><td>Question 10</td><td colspan="2">For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]</td></tr>
<tr><td>Consensus</td><td>Overview of assessments before the discussion:

Colour Code: green: negative; red: positive; orange: weakly active; yellow: equivocal</td><td>Overview after discussion:

 
</td></tr>
</table>

	There was some discussion on the role of the luciferase assay: it is not very indicative of a non-specific interference with TPO. Luciferase is indicative of quenching but quenching is not specific to TPO. Some assessors decided to ignore luciferase activity in their assessment of TPO inhibition of tested chemicals. H083: looks positive but there is evidence of a general degradation of the protein I488: inhibitor but not as clear as some of the other chemicals L465: dismiss the luciferase assay, everyone agreed that this is a weakly positive M192: Many runs were performed before getting the 20% inhibition, only 2 runs reach 20% inhibition and the baseline level of the enzyme activity is negative, hence the equivocal outcome. U778: positive as a consensus V050: consensus on a positive, just showed a problematic negative baseline activity. U796: results are difficult to interpret (incl. a negative baseline activity), consensus to call it an equivocal outcome. AD060: consensus that it is a clear inhibitor, but rated 'weak' because of the concentration at which the effect is elicited; AF 364: consensus to call it an equivocal; results are very messy and difficult to interpret.

After the assessment of Part 2 the JRC, the EU-NETVAL laboratory RISE and Method Developer US-EPA have met to discuss the questions that the assessment group had about the results generated with the AUR-TPO method. The summary of that meeting and some additional data that were generated by RISE during the validation study were shared with the assessment group.

The method developer and EU-NETVAL laboratory believed that the main issue that caused variability in the data has been the decreasing activity of TPO and Amplex UltraRed during the performance of the method. Recommendations for how to update the SOP have been provided, so that this can be avoided in future studies.

Part 2 report when the chemicals are unblinded:

Question 11	Now that the identity of the chemicals is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	It was agreed that only information from similar assays should be taken into account and that the results should not be compared with those from the tyrosine iodination assay. That changed the opinion from one assessor for several chemicals. In general, the results for most chemicals were in line with the expected results. The positive results of silicristin, tetrabromobisphenol A and pentachlorophenol, for which no data existed elsewhere, could be explained by their chemical properties. Tetra Methyl Thio Urea and Ethylene Thio Urea were not as positive as expected, but for these chemicals not all results are consistent in the literature and depend on the assay type and TPO source. In another experimental setup that includes the presence of iodine, these chemicals are expected to be more active.
Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	The hypothesis is that there were some issues with the degradation of the enzyme between the plates. The baseline was variable and sometimes very low. One suggestion is to change the SOPs (scraping the cells rather than using trypsin to have an optimum enzyme activity, reducing the time between enzyme and AUR preparation and exposure and measurement of activity, changing from single pipette to higher throughput device). But still there was some doubt that this SOP optimisation would solve all issues identified with the results.
Question 13	How would you judge the specificity of the method?
Consensus	For chemicals for which data exist, the specificity seems good. For many chemicals, there is no reference data to compare with, which makes specificity difficult to calculate.
Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	100 µM is reasonable. For soluble chemicals (drugs), there might be reasons to test at higher concentrations.
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	Once the SOPs are optimised, it may be possible to achieve 3 valid runs meeting the acceptance criteria.
Question 16	What is the group conclusion on the validation status of the assay?

Consensus	The assay is not yet ready and needs optimisation of the SOPs and re-testing to see if the SOPs improve the outcome of the assay or if there are other issues to address. A more automated approach (96 well plates or robot) would seem more appropriate for this assay.
Question 17	What further work (if considered necessary) would the assessment group recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method or to confirm the acceptance criteria or further development of the data interpretation procedure?)
Consensus	The enzyme preparation (including scraping the cells from the dish rather than using trypsin to avoid early enzyme degradation) should be improved, the maximum pipetting time should be specified, and the baseline response should then be checked. The same lab may need to run the assay again with optimised SOPs using the following 5 chemicals: MMI, ETU, rosmarinic acid, triclosan, and 2,4,6-Tribromophenol to confirm that the issue with the SOPs and the operator factor is settled. The assessment group mentioned that these chemicals do not need to be tested blind. In a second step, it would also be good to transfer the SOP to another lab that has a higher throughput capacity (e.g. robot or 96-well pipetter system). It was not discussed how many chemicals should be included in a transferability phase. Reliability and sensitivity of the assay need further work before transferability.

After completion of the assessment, the EU-NETVAL laboratory RISE performed an additional study with 5 chemicals recommended by the assessment group. This study should provide more reproducible data.

ANNEX 3- TYROSINE IODINATION USING LIQUID CHROMATOGRAPHY (TYRO-IOD)

NON-blinded phase with reference and control chemicals (Part 1)

TYRO-IOD Part 1 report: 5 runs are available for the few chemicals tested, including the reference and control items of the method.

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent.
Question 2	Is the reproducibility across the 5 runs consistent for the 7 chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	The reproducibility across the runs is good for all chemicals.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	Very good. Presence of a few outliers; when outliers are taken out the reproducibility is very good. Method used to identify outliers would be good to know as some outliers are not so obvious.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent.
Question 5	Any other observations on the method
Consensus	----------
Question 6	Any other observations on the data
Consensus	The assessment group is interested to know more about methods to determine the z factor across different experiments, variability (CV, SD) of the reference chemicals.

Conclusion from the assessment group on Part 1 report?
Recommendations from the assessment group for further standardisation of method?

- Method used to identify outliers would be good to know as some outliers are not so obvious.
- Question for clarification to the lab: how did they manage with the cell batches?
- Question on the throughput level to be asked to the lab? Medium throughput is 96 well-plate.

After the meeting the EU-NETVAL laboratory clarified that:

- There was no specific test used for outlier removal. Outliers were removed based on the overall curve profile and/or the corresponding replicates. This comment should have been added to all relevant data tables.
- All incubations were performed in U-tubes. The U-tubes can be used in a 96-well format.
- One batch of the TPO cell lysate was used for the whole study.

Part 2 report when the chemicals are blinded:

Part 2 report: 3 runs are available for 30 chemicals tested, with 8 concentrations/chemical.

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Comment from JRC: "In most cases, a third valid run was not needed because two concordant valid runs were sufficient in most cases to conclude". Overall the reproducibility across runs was excellent. Reproducibility is good. The group wonders how would reproducibility be, should different batches have been used rather than the same batch between Part 1 and Part 2. The Group recommends to ask the lab whether they have done any batch-to-batch comparison before. Overall the level of standardisation seems very good (as judged by MIT concentration).
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	Generally consistent. Small differences noted but with no impact on the conclusions. Very good fit and close effective values, despite the differences in concentrations tested.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	"Active" at a certain percentage, different thresholds, one threshold around 20%, another threshold around 50%. Needs to be different from the noise/background. Beyond 20% inhibition is called "active" Some indication of potency would be helpful: Active/stimulating/equivocal/weak or inactive. If a chemical would reach an IC50, the IC20 should also be reported, not to lose important information. If the IC50 is not reached because of various reasons (solubility issue for example or IC50 higher than the max tested concentration), report the IC20 (and the IC50 as being higher than the max tested concentration).

Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	

	Chemical code	KH	OB	SD	AF
1	563	WEAK	WEAK/CLEA	POS	POS
2	431	WEAK	WEAK	POS	POS
3	140	POS	POS	POS	POS
4	597	WEAK	WEAK	POS	POS
5	226	POS	POS	POS	POS
6	301	POS	POS	POS	POS
7	550	NEG	NEG	NEG	NEG
8	515	POS	POS	POS	POS
9	254	POS	POS	POS	POS
10	661	NEG	NEG	NEG	NEG
11	920	NEG	NEG	NEG	NEG
12	532	WEAK	WEAK	POS	POS
13	401	POS	POS	POS	POS
14	310	WEAK	WEAK	POS	POS
15	482	POS	POS	POS	POS
16	696	POS	POS	POS	POS
17	136	NEG	NEG	NEG	STIM
18	779	WEAK	WEAK	POS	POS
19	135	WEAK/EQU	WEAK/EQUI	POS	NEG
20	834	NEG	NEG	NEG	NEG
21	303	POS	POS	POS	POS
22	387	WEAK	WEAK	POS	EQUIV
23	071	POS	POS	POS	POS
24	151	WEAK	WEAK	POS	POS
25	820	WEAK	WEAK/EQUI	POS	EQUIV
26	012	WEAK	WEAK/EQUI	POS	EQUIV
27	246	NEG	NEG	NEG	NEG
28	341	POS	POS	POS	POS
29	805	POS	POS	POS	POS
30	992	POS	POS	POS	POS

Part 2 report when the chemicals are unblinded:

Question 11	Now that the identity of the chemicals is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	With this question also the responses provided to question 10 (data interpretation and assignment of clear/weak/equivocal/negative result for the chemicals) were discussed. The differences in the judgement was explained by the use of different criteria by the experts. For other evaluations it was recommended that these criteria are better defined. There was generally good concordance between the expected result (based on chemicals selection), and the response obtained in the assay. For about half of the chemicals the positive (twelve chemicals) or negative (four chemicals) outcome was known, and confirmed with the method. There were no known inhibitors resulting in negative outcome in the TYRO-IOD assay, indicating a good sensitivity. For the other half of the chemicals tested, there were a number of chemicals with actually limited knowledge on their TPO inhibition, that turned positive (mostly weak) in the TYRO-IOD assay. There was generally the impression from the assessment group that some of the chemicals were expected to be negative in the assay on the basis of 'no knowledge' of their TPO inhibitory potential rather than documented evidence of their non-activity in this assay, while they are known positives on other thyroid modes of action. For the supposedly negative chemicals that resulted in positive outcomes in the TYRO-IOD assay, individuals in the group had difficulty finding information in the literature, or in ToxCast or other database on the (lack of) TPO activity. The group would be interested to corroborate data for those expected negative chemicals with data obtained in the TPO-AUR assay. There were a few surprises though such as ampicillin, which is supposed to be negative for all thyroid *in vitro* methods; this might be explained by some non-specific activity, although previous similar *in vitro* testing using rat microsomes, and *in vivo* tests resulted in negative outcomes. [*Note from expert: Provided that the cited assays did not use iodide as TPO substrate, the negative outcome is logical. In a chemical setup, penicillins can be oxidized by iodine in solution. In the presence of TPO/iodide/hydrogen peroxide the generated iodinating species (hypoiodous acid?) is the corresponding oxidant.]* A suggestion was made to consider some threshold for the IC50, e.g. 50µM, for an outcome to be considered positive in the assay, but that was not discussed in detail.
Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	The assessment group agreed that there is excellent consistency across runs and the question is not really relevant in the case of this dataset.
Question 13	How would you judge the specificity of the method?
Consensus	The group was in agreement that for chemicals that are documented negatives (from multiple sources including ToxCast) there is good concordance with the negatives found in the TYRO-IOD assay. For other chemicals that are supposed negatives (in the sense that there are no reliable data indicating the substance as either active or

	inactive for this mode of action), not much could be concluded in the absence of sufficient data.
Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	Generally 100 µM is the consensus top concentration. In a TG, it should be the default max concentration, unless there is ADME information or a specific toxicological context to justify testing higher concentration(s).
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	There was agreement that 3 runs are enough. In a strategy to limit testing to the minimum necessary, 2 negative runs could be sufficient to conclude on the inhibition of TPO activity. In case of discordant runs or in case of 2 consecutive positive runs, a third run could/should be performed, possibly within a narrower range of concentrations to clarify/confirm the result.
Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The group was in general agreement that the results obtained in the laboratory are "clean" and reproducible (concordant runs and reproducible dose-response curves), the assay is sensitive (all positives correctly identified) and specific at least for the documented negatives. The remaining question that has not been addressed is the transferability of the assay. Indeed, since the laboratory developed their own standard operating procedures (thus acting in part as a method developer), the reproducibility of results on selected chemicals in at least one or two other laboratories would be important to confirm the assay is robust when performed outside the method developer's testing facility.
Question 17	What further work (if considered necessary) would the assessment group recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method or to confirm the acceptance criteria or further development of the data interpretation procedure?)
Consensus	Some in the group advocated that two other laboratories are necessary to evaluate assay transferability, while another member of the group thought that one laboratory could be enough, provided some chemicals are tested under blind conditions; adding a second or a third laboratory does not change the assay, just the perception/confidence we have in the assay. Generally the whole assessment group agreed it would be good to have blind testing systematically included in the transferability phase, with focus on the proficiency chemicals for the TG. Laboratories would need both cell culture facilities and the analytical instrumentation needed. There was general agreement that a limited number of well characterised chemicals (2-3 positives/2-3 negatives) might be enough; the 5 chemicals from Part 1 could be used or a sub-set of the 30 chemicals from Part 2 (see below). It was also suggested to consider in that list thioureas, for the purpose of comparison with the TPO-AUR assay. Also, a comparison of results obtained in Part 2 with the TPO-AUR results would be useful to corroborate findings and develop the data interpretation procedure. According to the experts there is no need to run a specificity test for non-specific

	inhibitors. So far only surfactants are known to interfere. During the review of Part 2 results, a set of chemicals were identified as suitable for further validation purposes.

ANNEX 4- THYROXINE-BINDING PREALBUMIN (TTR) / THYROXINEBINDING PREALBUMIN (TBG) BINDING USING FLUORESCENCE DISPLACEMENT (ANSA). (TTR-ANSA)

NON-blinded phase with reference and control chemicals (Part 1)

TTR ANSA Part 1 report: 5 runs are available for few chemicals tested mostly including the reference and control items of the method.

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	Good reproducibility for TTR and fair to poor for TBG.
Question 2	Is the reproducibility across the 5 runs consistent for the 7 chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	For TTR fairly consistent to consistent (higher variability for BPA and TBBPA), for TBG, many invalid experiments and inconsistencies.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	Based on the CV and error bars (presuming they represent variability between replicates) for the IC50, very good to excellent variability for TTR. Variability for TBG was poor (high variability).
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	When looking at the raw data, the dynamic range is within a factor of 3 between background noise and signal, which is limited although this range is relatively low, it is fairly good for such fluorescence enhancement assays.
Question 5	Any other observations on the method
Consensus	Not very clear how the data correction was made using the controls (positive control was not a true positive control). This part of the protocol and data interpretation may need further discussion. One assessor's considerable work on the TBG assay (high protein purity) showed that the assay performs well in his lab (publication to come soon). For compounds that are fluorescent, this type of assay does not work well and this should be considered upfront as a limitation of the assay.
Question 6	Any other observations on the data
Consensus	It was generally difficult to find the dynamic range of the assay.

Part 2 report when the chemicals are blinded:

Part 2 report: 3 runs are available for 30 chemicals tested, with 8 concentrations/chemical.

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent. Assessors noted discrepancies between tables and corresponding curves. For clear inhibitors, there was good overlap across the runs but for other chemicals, there were discrepancies noted and there will be a quality check performed on the data, and subsequent corrections made before the unblinding of the chemicals.
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	Yes, good reproducibility especially for clear inhibitors, more variability for low responses. Data in the tables seem to show more variability than the curves for some chemicals. Quality check will be performed and corrections made.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	Both the IC20 and IC50 should be reported, when reached (specify whether measured or extrapolated). A chemical is (or may be) considered active if a statistically significant displacement beyond 20% is measured, in a concentration-dependent way, at a maximum of 200 micromolar. A 20% displacement that is not statistically significant but still concentration-dependent should be reported anyway.

Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	There was consensus on all of the chemicals assessed and a note that the assessment was made in the absence of knowledge on autofluorescence, which makes it difficult to judge activity.

Blind Part 2 assessment for TTR ANSA, before discussion

Chemical code	KH	OB	TH	SD
10				
186				
239				
338				
340				
386				
407				
463				
510				
545				
599				
673				
740				
755				
889				
13				
37				
72				
164				
214				
266				
283				
299				
481				
501				
516				
698				
724				
743				
831				

Blind Part 2 assessment for TTR ANSA, after discussion

Chemical code	KH	OB	TH	SD
10				
186				
239				
338				
340				
386				
407				
463				
510				
545				
599				
673				
740				
755				
889				
13				
37				
72				
164				
214				
266				
283				
299				
481				
501				
516				
698				
724				
743				
831				

Part 2 report when the chemicals are unblinded:

Question 11	Now that the identity of the chemicals is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	Generally the results achieved were conform the expectations. Possible auto-fluorescence or quenching was not tested in Part 2 and could be a source of interference. High concentrations may lead to solubility issues, possibly leading to interferences as well.
Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	*Note from JRC: there were some mistakes on the curves/data analysis from the laboratory that became apparent between the blind evaluation and the unblinded evaluation of Part 2. Corrections were made.* After the correction, the consensus is that there is good consistency across runs for all chemicals except sobetirome where there might still be a mistake on the curves/data analysis (?) but no explanation could be found.
Question 13	How would you judge the specificity of the method?
Consensus	Good specificity. When comparing with the other assay (TTR-FITC) some of the negative chemicals in TTR-ANSA were judged as weakly positive with TTR-FITC because they were tested at higher concentrations in the TTR-FITC. A remark was made on the high DMSO concentration which would increase background fluorescence, something to further investigate to improve the dynamic range of the assay.
Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	For the majority of environmental chemicals, 100-300 uM seems to be the consensus for a technically feasible/reasonable maximum concentration; exceptionally and for the sake of having a more complete dose-response curve for highly soluble chemicals, there may be utility testing as high as 1000uM in case there is some indication of effect at 100uM. However, there was some discussion around the limited value and possible interpretation of results (positive/negative) beyond 300 uM. This discussion is not specific to this assay and merits further general discussion and guidance.
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	3 runs is a good number.
Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The assay works, is sensitive and specific and reproducible (also when results are

	compared with the US EPA Duluth data, 6 overlapping chemicals with full concentration-response curves, 11 overlapping chemicals with few data points). Some housekeeping/cleaning needs to be operated on the SOP: to address possible interference with fluorescence readout, better describe approach for calculations and data analysis, the composition/definition of positive control and negative control need to be revised (see response to Q5); the solubility interference with the fluorescence readout.
Question 17	What further work (if considered necessary) the assessment group would recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method? or to confirm the acceptance criteria? Further development of the data interpretation procedure?)
Consensus	For the SOP: further work needed on the criteria for interference with the fluorescence readout. Once the SOP has been updated based on the response to Q16 above, further transferability to at least one, better two, labs would be needed. A sub-set of Part 2 chemicals would need to be selected and blind tested, another call of the assessment group would be needed to propose candidate positive/negative chemicals for the transferability.

ANNEX 5- THYROXINE-BINDING PREALBUMIN BINDING USING FLUORESCENCE DISPLACEMENT. (TTR FITC T4)

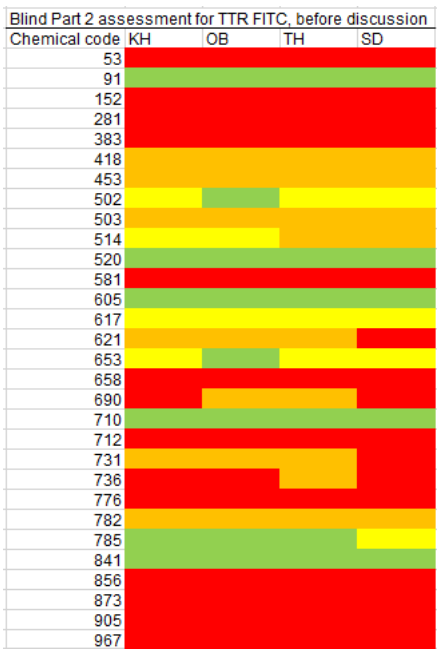
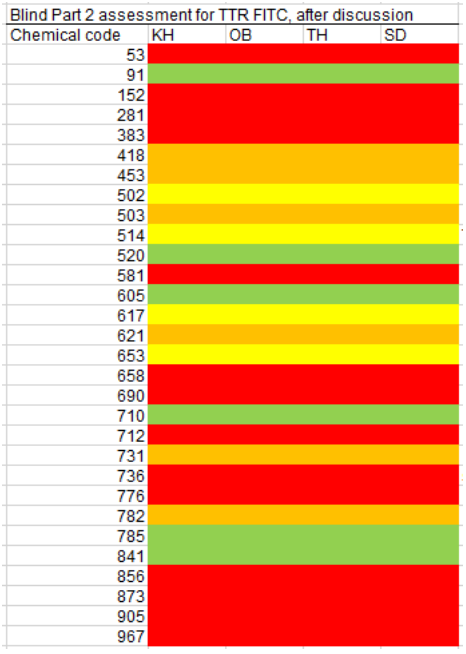
TTR FITC NON-blinded phase with reference and control chemicals (Part 1)

TTR FITC Part 1 report: 5 runs are available for few chemicals tested mostly including the reference and control items of the method.

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent.
Question 2	Is the reproducibility across the 5 runs consistent for the 7 chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	Yes, generally consistent for all chemicals, except for TBBPA (curve is very steep with limited data points).
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	When looking at the raw data, the dynamic range is within a factor of ~2.2 between background noise and signal, which is limited although this range is relatively low, it is fairly good for such fluorescence enhancement assays.
Question 5	Any other observations on the method
Consensus	In the method description, there should be an explanation of how the slope of the autofluorescence acceptance criteria is derived (on what basis, i.e. autofluorescence of the compound tested) and how to adjust the acceptance range for lab- and equipment-dependence. Out of the acceptability range, the data of autoflurorescent compounds can still be used as an indication of binding and another assay (radioactive immune assay) should be used for confirmation.
Question 6	Any other observations on the data
Consensus	It was generally difficult to find the dynamic range of the assay, it would be good to provide the dynamic range and the Z factor with the report.

Part 2 report when the chemicals are blinded:

Part 2 report: 3 runs are available for 30 chemicals tested, with 8 concentrations/chemical.

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent.
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	Generally YES. There are a few exceptions with higher variability for example for chemicals with weak activity.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	Same as for ANSA.
Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	 

Part 2 report when the chemicals are unblinded:

Question 11	Now that the identity of the chemicals is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	Generally the results achieved were conform the expectations. Test concentrations were very high for some chemicals, which resulted in a weakly active judgement.
Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	The concentration response curves for dibutylphthalate were variable across runs, with U-shape at very high tested concentrations. Possible solubility issues (precipitation?) and interference with the fluorescence may be the reason. All the other chemicals were consistent between the runs.
Question 13	How would you judge the specificity of the method?
Consensus	Good specificity. When comparing with the other assay (TTR-ANSA) some of the negative chemicals in TTR-ANSA were judged as weakly positive with TTR-FITC because they were tested at higher concentrations in the TTR-FITC.
Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	For the majority of environmental chemicals, 100-300 uM seems to be the consensus for a technically feasible/reasonable maximum concentration; exceptionally and for the sake of having a more complete dose-response curve for highly soluble chemicals, there may be utility testing as high as 1000uM in case there is some indication of effect at 100uM. However, there was some discussion around the limited value and possible interpretation of results (positive/negative) beyond 300 uM. This discussion is not specific to this assay and merits further general discussion and guidance.
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	3 runs is a good number.
Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The assay works, it is sensitive and specific and reproducible (also when results are compared with the VU Amsterdam's lab on 11 chemicals). The acceptability criteria for the slope used for the assessment of the interference with fluorescence readout needs to be defined more clearly.
Question 17	What further work (if considered necessary) the assessment group would recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method? or to confirm the

	acceptance criteria? Further development of the data interpretation procedure?)
Consensus	For the SOP: further work needed on the criteria for interference with the fluorescence readout. Generally the assessment group thinks the SOP for this assay is more ready than the SOP of the other TTR-ANSA assay. Further transferability to at least one, better two, labs would be needed. A sub-set of Part 2 chemicals would need to be selected and blind tested, another call of the assessment group would be needed to propose candidate positive/negative chemicals for the transferability.

ANNEX 6- DEIODINASE 1 ACTIVITY BASED ON SANDELL-KOLTHOFF REACTION.

NON-blinded phase with reference and control chemicals (Part 1)

DIO1 Part 1 report: 5 runs are available for few chemicals tested mostly including the reference and control items of the method.

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	Pretty good reproducibility, considering this is early transfer of a method. Small variations noted among chemicals, regarded as “normal” and possibly attributed to variations in the execution of the protocol more than assay robustness. Variation has relatively limited impact on the interpretation of data.
Question 2	Is the reproducibility across the 5 runs consistent for the 7 chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	Yes, there are differences, e.g. genistein, for which solubility issues arise. Also TBBPA suffers from some variability across the 5 runs and further optimisation and concentration range finding are expected to solve solubility, and variability issues. Incomplete dose-response curves is also impeding a good comparison between runs. Even though there were solubility issues, they have at least been recognised and the IC50 is ultimately quite similar with the literature.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	Excellent, based on the z factor (difference between the max response in the positive control (6-PTU) and the control). Question was initially not understood in the same way

	by all assessors.
Question 5	Any other observations on the method
Consensus	The use of human microsomes raised some issues for further consideration: 1) viral load and ways to check and reduce it, 2) variability from batch to batch and possibility of interaction with test chemical (in case of negative result, is it truly negative or did a CYP enzyme deactivated the test chemical?). Batch variability can be addressed by large pools of donors. 3) possibility to use alternative sources of enzymes (liver extract from rat or mouse, recombinant source,…) but leads to major differences in assay outcome due to different metabolic interactions with the test chemical. Consideration to have both sources (pure and mixed enzyme) in the same assay. Enzymatic source should be commercially available, which may not be the case of the recombinant source. Recommendation from the assessment group: continue the evaluation and try to find solutions for the human microsome issues identified. Perhaps we could recommend that the assay platform, based on the current results, is reliable (pending results from broader chemical space). We could also recommend that an additional model (recombinant enzyme or from human cell lines without much drug metabolism capacity) might also be considered.
Question 6	Any other observations on the data
Consensus	The lab has not been using a 96-well plate pipetting device. This could be a recommendation for the future to improve consistency.

Conclusion from the assessment group on Part 1 report?

- [Nothing specific was mentioned by the assessment group]

Recommendations from the assessment group for further standardisation of method?

- [Nothing specific was mentioned by the assessment group]

Part 2 report when the chemicals are blinded:

Part 2 report: 3 runs are available for 40 chemicals tested, with 8 concentrations/chemical.

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Good to Excellent. Compounds that are less active result in more variability. Overall the group was positively surprised at how reproducible the assay is across the 40 chemicals tested, considering the human microsomes are used (which was a source of questioning during Part 1 assessment).
Question 8	Is the reproducibility across the runs consistent for the chemicals tested?

	In the case you see an issue with one chemical, please flag it.
Consensus	Reproducibility across the runs is very good for all chemicals. There seems to be no correlation between the % response (inhibition) and the % variation. The group noted that sometimes the full curve is not available for weakly active substances (no IC50) so there could be another way to analyse variability, e.g. using the PC10 or PC20.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	Considerations from the group around a positive call in the assay: • 25% inhibition • determination of IC50 at the highest concentration or two consecutive concentrations, within solubility limits; • lack of activity in a microsome-free sample (rule out false positive due to unspecific activity, as a cytotoxicity surrogate) There was discussion around the fact that the DIO1 assay is particularly vulnerable to the SK reaction, consequently one should be careful not over-interpreting this assay in isolation. The group also discussed interference with ALP and whether to integrate it in the DIP but there was no consensus, rather preference to keep ALP interference as a piece of information to consider when evaluating activity across assays in an IATA context.
Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	There were discussions over how to judge an equivocal result: there are various reasons why a result can be equivocal. One assessor mentioned that potency cannot be judged by the % inhibition only. Can be because an assay is only positive at highest concentration tested, but then the maximum concentration tested needs to be taken into account (too low?). Also the solubility issue needs to be considered in such case.

Part 2 report when the chemicals are unblinded:

<table>
<tr><td>Question 11</td><td colspan="2">Now that the identity of the chemicals is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?</td></tr>
<tr><td>Consensus</td><td colspan="2">Before discussing specific chemicals, the group went through the calls and re-discussed the activities consensus on basis of the shapes of the dose-response curves, especially because JBF and MJ had changed some calls after the last meeting.

It was agreed that inhibition of alkaline phosphatase (ALP) is not indicative for the inhibition of DIO1, so when a chemical is both inhibiting ALP and DIO1, it is considered a positive.

There were a few adjustments considering that ALP interference is non-specific and thus some equivocal calls were transformed into (weak) positives. Some assessors did not discriminate between weak and clear or strong positives, and each used their own criteria to discriminate among positives.

There was a question from one assessor on the (positive) results from linoleic (818) and linolenic (160) acids. The esterified form of those acids is present in the diet. but only the free fatty acid is known to have a surfactant effect resulting in positive outcome in the DIO1 assay. It should be kept in mind that the positive results for linoleic acid are not indicative for the derivative used in food.

The ALP interference raised the need to develop a tiered approach to data interpretation, i.e. ALP interference should not be used in the interpretation of one particular assay, but rather at the level of a defined approach using multiple data sources.

There was recognition that it may be useful to generate potency information from the data generated, when possible. One way to do that in an objective manner would be to compare to the reference chemical used in the assay (here PTU). Such comparison could be done with complete dose-response curves, or via modelling of the bench mark dose (BMD), in case the curve does not reach 50% inhibition.

The experts agreed there was a need for another meeting to confirm if the results from these chemicals are as expected.</td></tr>
<tr><td></td><td>Before unblinding
(initial individual assessment)</td><td>Before unblinding (final interpretation on the basis of group discussion, not considering chemical identity, while disregarding ALP results)</td></tr>
</table>

Chemical code	JBF	KR	MJ	SD
56	POS	POS	POS	POS
125	EQUIV	WEAK	POS	POS
130	NEG	NEG	NEG	NEG
160	EQUIV	WEAK	EQUIV	POS
194	EQUIV	WEAK	NEG	POS
218	NEG	NEG	NEG	NEG
220	WEAK	WEAK	WEAK/EQ	POS
227	POS	POS	POS	POS
229	EQUIV	POS	EQUIV	POS
279	POS	WEAK	WEAK	POS
294	POS	WEAK	WEAK	POS
307	POS	POS	POS	POS
325	POS	WEAK	WEAK	POS
377	NEG	NEG	NEG	NEG
437	NEG	NEG	NEG	NEG
442	NEG	NEG	NEG/EQU	NEG
506	WEAK	WEAK	WEAK	POS
511	NEG	NEG	NEG	NEG
526	POS	WEAK	WEAK	POS
543	POS	POS	POS	POS
551	NEG	NEG	NEG	NEG
598	NEG	NEG	NEG	NEG
603	EQUIV	EQUIV	EQUIV	EQUIV
610	POS	POS	POS	POS
615	EQUIV	EQUIV	EQUIV	EQUIV
667	EQUIV	WEAK	EQUIV	POS
680	WEAK	WEAK	POS	POS
741	POS	WEAK	WEAK	POS
798	POS	WEAK	POS	POS
818	POS	WEAK	POS	POS
827	NEG	NEG	NEG	NEG
839	POS	WEAK	WEAK/EQ	POS
850	WEAK	WEAK	POS	POS
868	NEG	NEG	NEG	NEG
877	NEG	NEG	NEG	NEG
878	POS	WEAK	WEAK	POS
925	NEG	NEG	NEG	NEG
933	POS	WEAK	WEAK	POS
940	EQUIV	EQUIV	EQUIV	EQUIV
974	NEG	NEG	NEG	NEG

Chemical code	JBF	KR	MJ	SD
56	POS	POS	POS	POS
125	WEAK	WEAK	WEAK	WEAK
130	NEG	NEG	NEG	NEG
160	EQUIV	EQUIV	EQUIV	EQUIV
194	EQUIV	WEAK	EQUIV	POS
218	NEG	NEG	NEG	NEG
220	WEAK	WEAK	WEAK/EQ	POS
227	POS	POS	POS	POS
229	POS	POS	POS	POS
279	POS	WEAK	POS	POS
294	POS	WEAK	WEAK	POS
307	POS	POS	POS	POS
325	POS	WEAK	WEAK	POS
377	NEG	NEG	NEG	NEG
437	NEG	NEG	NEG	NEG
442	NEG	NEG	NEG/EQU	NEG
506	WEAK	WEAK	WEAK	POS
511	NEG	NEG	NEG	NEG
526	POS	WEAK	POS	POS
543	POS	POS	POS	POS
551	NEG	NEG	NEG	NEG
598	NEG	NEG	NEG	NEG
603	EQUIV	EQUIV	EQUIV	EQUIV
610	POS	POS	POS	POS
615	EQUIV	EQUIV	EQUIV	EQUIV
667	EQUIV	WEAK	EQUIV	POS
680	WEAK	WEAK	WEAK	POS
741	POS	WEAK	POS	POS
798	POS	WEAK	POS	POS
818	POS	WEAK	POS	POS
827	NEG	NEG	NEG	NEG
839	POS	WEAK	WEAK/EQ	POS
850	WEAK	WEAK	POS	POS
868	NEG	NEG	NEG	NEG
877	NEG	NEG	NEG	NEG
878	POS	WEAK	WEAK	POS
925	NEG	NEG	NEG	NEG
933	POS	WEAK	WEAK	POS
940	EQUIV	EQUIV	EQUIV	EQUIV
974	NEG	NEG	NEG	NEG

Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	When there were inconsistencies across the runs, the assay itself did not appear to be the issue, rather the solubility or the stability of the tested chemical could explain the discrepancies (e.g. isoflavone, genistein), due to sensitivity to light or temperature.
Question 13	How would you judge the specificity of the method?
Consensus	The group consensus is that the assay has a good specificity.
Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	The group generally agreed that testing above 1mM should not be done as it would be disruptive to the test system and perturb the toxicokinetics (e.g. by engaging reaction with the co-factor rather than the enzyme).
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	The group agreed that 3 runs are sufficient. A fourth run might be considered in cases where there is >20% variation between the runs.

Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The group concurred that the validation status (intra-laboratory reproducibility of the assay) is good.
Question 17	What further work (if considered necessary) would the assessment group recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method or to confirm the acceptance criteria or further development of the data interpretation procedure?)
Consensus	The group proposed that laboratory transferability should be evaluated by transferring the assay to at least another laboratory, testing the reference chemicals from Part 1 (2 positives and 1 negative) and a sub-set of the 30 chemicals from Part 2 in a blind fashion. Candidate chemicals (extracted from Excel sheet): (not disclosed here in this report in case of blind testing). Although no clear consensus was achieved, proposals were to blind test among the chemicals selected: - 2-5 clear positive, - 2-5 clear negatives, - 2-5 weak positives.

ANNEX 7- INHIBITION OF THYROID HORMONES (THS) GLUCURONIDATION USING LIQUID CHROMATOGRAPHY/MASS SPECTROMETRY (LC/MS)

NON-blinded phase with reference and control chemicals (Part 1)

GLUC—INH-LCMS Part 1 report: 6 runs are available for few chemicals tested mostly including the reference and control items of the method.

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent (one out of 5 runs probably had a technical issue (missed the addition of Alamecin?), to be confirmed). Both activity and inhibition could/should be evaluated (although they are presumably linked). Mostly the inhibition was assessed under Q1. For the activity assay, it will be checked and confirmed whether the acceptance criteria requires a certain activity level of the microsomal fraction. From Part 1 report, there seems to be no problem with the activity of the enzyme. Following the first call and additional views provided on the MU fluorescence assay, the consensus is that reproducibility is good across the 5 runs.
Question 2	Is the reproducibility across the 5 runs consistent for the chemical tested? In case you see an issue with the chemical, can you identify/hypothesise the reason?
Consensus	It should be revisited for Part 2 report as here for Part 1 there is only one test chemical and it actually meets the expectation of a positive chemical. It was clarified that the lab tested for the solubility and there was no problem.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	Good to excellent, considering the acceptable variation should remain within 20%.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	Generally good, but there were some questions within the group on how to interpret the assay and the dynamic range. It seems that it is generally superfluous to test 3 concentrations of the substrate, as 10 microM seems to be enough to test and does not generate so much variability. It was confirmed that in Part 2, only 10 microM was used. The dynamic range is understood as the percentage of inhibition achieved within the range of concentrations tested (1.5 to 2 log). However, this assessment is more a reflection of the responsiveness of the assay, rather than a distinction of the maximum signal (i.e. maximum fold activity) in the control compared to the background noise. In this context, the use of 10 microM substrate improves the dynamic range as the activity level starts above the control; above 10 microM there seems to be solubility issues (precipitation).

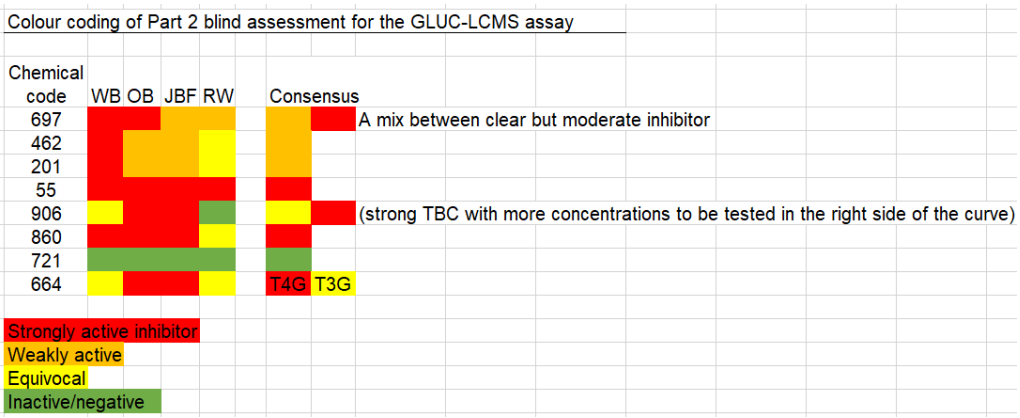
Question 5	Any other observations on the method
Consensus	---
Question 6	Any other observations on the data
Consensus	---

Part 2 report when the chemicals are blinded:

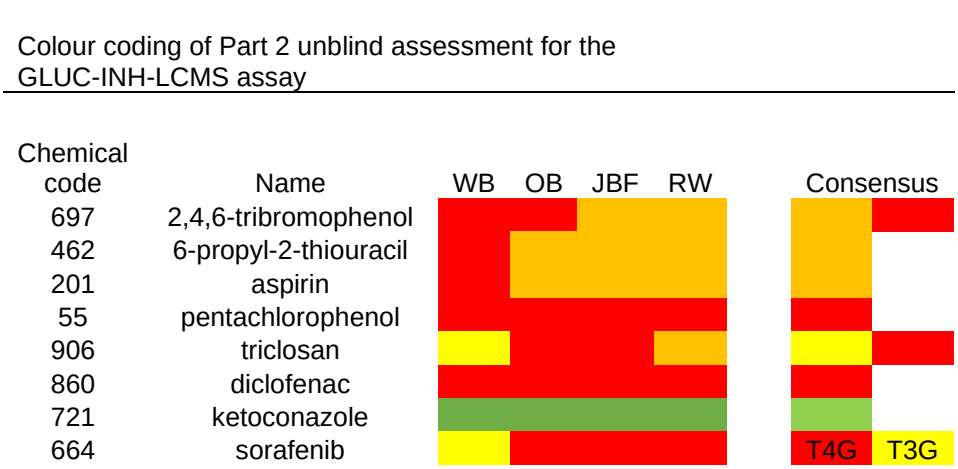
Part 2 report: 2 runs are available for 8 chemicals tested, with 4-6 concentrations/chemical. Each run was performed in triplicate.

[It was clarified that the runs were done in triplicate. The coefficient of variation was calculated on the basis of the variation within run].

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]? [This question can only be answered when more than 1 run is provided]
Consensus	Good reproducibility between the 2 runs (for T3 and T4).
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	[It was requested to clarify what data exactly is behind each run; which run was accepted for each chemical]. From a first glance, reproducibility seems to be good, but the group will reconsider when the data used to calculate coefficients of variation has been clarified.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	The group would set the criteria around 20% (maximum 30%) for discriminating a positive from a negative. In relation to potency, there was discussion on the best way to express it, and what criteria to use (IC50 or IC20 achieved at [given conc.]? Or maybe relative to reference positive chemical?). There was also discussion on the biological relevance (i.e. meaning) of the qualifiers “weak” and “strong”, taken in an *in vivo* situation, compared to the context of an *in vitro* assay.

Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	

Part 2 report when the chemicals are unblinded:

Question 11	Now that chemicals identity is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	
Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	Inconsistency issues between runs is not frequent.
Question 13	How would you judge the specificity of the method?

Consensus	Given the limited number of negative chemicals tested, it is difficult to judge the specificity of the method at this stage. It would be appropriate to test a few more negative chemicals.
Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	The assessment group had difficulties providing a given concentration. It probably makes sense to test until water solubility, provided that results generated from the dose response curve are interpreted in the context of the toxicokinetics and weight of evidence. A dose range finder should be performed to avoid generating data that are difficult to interpret (e.g. very steep curve between the top doses).
Question 15	Are 2 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	Two runs are sufficient if they are consistent, if not consistent a third run would be needed.

Conclusion and recommendation from the assessment group:

Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The group generally agreed that the assay seems to perform well (good reproducibility). The evaluation of specificity would benefit from testing additional negatives. However, while the Expert Group on Thyroid Disruption Methods (TDM EG) generally agreed that the assay performs as designed, they questioned the physiological relevance of the assay which measured glucuronidation inhibition whereas the most frequent effects found in vivo are TH reduction via increased glucuronidation.
Question 17	What further work (if considered necessary) the assessment group would recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method? or to confirm the acceptance criteria? Further development of the data interpretation procedure?)
Consensus	Further transferability of the assay to at least one other laboratory would be needed to judge reproducibility. Some additional review of the UGT metabolites within the assay and the role of specific enzymes on its performance would be helpful in interpreting data generated in some cases/for some chemicals. Further testing of triclosan at the right dose levels would be important to confirm its utility in further transferability studies.
Question 18	For which chemicals there is sufficient information that they are active or inactive for the mode of action? Please indicate those, so that they can be considered for follow-up validation studies.
Consensus	Current suggestions are: Pentachlorophenol, sorafenib and diclofenac as a positive (pan vs. T3/T4 specific UGT), ketoconazole as a negative. Further literature review would be needed to increase confidence that these chemicals are good reference

	chemicals for further work.

ANNEX 8- HUMAN THYROID HORMONE RECEPTOR ALPHA (TRα) AND HUMAN THYROID HORMONE RECEPTOR BETA (TRß) REPORTER GENE TRANSACTIVATION MEASURING AGONIST ACTIVITIES

NON-blinded phase with reference and control chemicals (Part 1)

Part 1: 5 runs are available for few chemicals tested mostly including the reference and control items of the method.

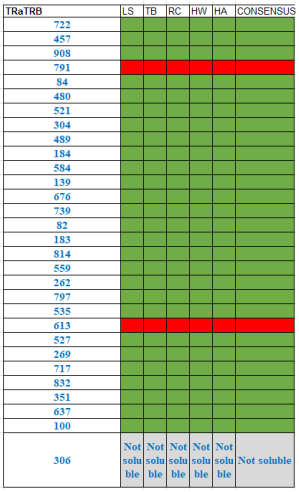
Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	The assessment group felt that the reproducibility across runs was better for TRbeta than for TRalpha. However, the lesser reproducibility across runs for TRalpha is not a show stopper.
Question 2	Is the reproducibility across the 5 runs consistent for the chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	The assessment group felt that there is consistency across the chemicals for the reproducibility between runs, following the trend that reproducibility is better for TRbeta than TRalpha. One hypothesis may be the lack of data normalisation for the% of viability (e.g. viability of 207%).
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	The group assessed variability within runs as fair to good after discussion and clarification.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	The dynamic range was evaluated as good to excellent (induction factor >1000).
Question 5	Any other observations on the method
Consensus	The group felt that all points raised above in response to question 5 were worth further attention by the test developer in further improving the assay. In relation to the HEK cell stability, an explanation was provided that the cells do not need to be cultured in the lab and the quality is assured by the cell provider (cell passage).
Question 6	Any other observations on the data
Consensus	The group would like to further reflect in the future, based on Part 2 data, if there would be an explanation for such a high induction factor compared to the TR-CALUX (is there any difference in receptor construct affecting the expression and induction factor? Any interference with luciferase activity and fluorescence measurements?). The group will further discuss in Part 2 whether this affects potentially the performance of the assay in general.

Part 2 report when the chemicals are blinded

Part 2 report: 3 runs are available for 2 over 30 chemicals tested (1 run for the negative chemicals), with 8 concentrations/chemical.

General remark:
One assay for agonist activity (incl. cytotox), no antagonism tested. 1 compound not soluble, 29 chemicals tested in dose range finding, with only 2 of 29 positive for both TRα and TRβ (TI 791 & TI 613). Only these 2 positive items were further tested in 3 runs, being sufficient as all were scored valid.

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Generally good (excellent in case the assay is used for hazard identification, fair if the data is used for risk assessment); it was noted though that there was an outlier in one run for chemical 791 for both assays (explanation provided by the lab was possibly a pipetting error), generating a concern in case the data would be used as such for decision making. An option in such case would be to run one or two more additional run to verify if this is a true outlier, and factor this in the decision criteria for the assay.
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	Generally see response to question 7. The low number of positive chemicals remains the limiting factor to evaluating the assay reproducibility.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	For hazard identification, two consecutive concentrations should show a signal with the highest signal above 20% activity and in the absence of cytotoxicity. If the assay had some potential use for risk assessment, an EC50 and EC20 should be determined from the concentrations tested.
Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]

Consensus	

TRaTRB	LS	TB	RC	HW	HA	CONSENSUS
722						
457						
908						
791						
84						
480						
521						
304						
489						
184						
584						
139						
676						
739						
82						
183						
814						
559						
262						
797						
535						
613						
527						
269						
717						
832						
351						
637						
100						
306	Not soluble	Not soluble	Not soluble	Not soluble	Not soluble	Not soluble

Part 2 report when the chemicals are unblinded:

Question 11	Now that the identity of the chemicals is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	There is a clear consensus that both sobetirome and TETRAC are clear positive agonists and all the other chemicals are negative. Discussion about PFOS, which has some binding affinity with the TR, but does probably not come out positive because of a different mode of action, i.e. not the classical binding-transactivation. There was also some discussion that the TR binding pocket is deep and thus chemicals do not compete well with T3 and T4 binding. There was some discussion about the TETRAC data that comes out positive (very high response in the viability assay measured by fluorescence in one run), but there was concern looking at the data that the result may be due to interference with fluorescence rather than cell viability, additional testing for interference with fluorescence would be needed.
Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	Not applicable in this case.
Question 13	How would you judge the specificity of the method?
Consensus	Good specificity, based on two positive chemicals.
Question 14	Looking at the available information (concentration-response curves (shape),

TRaTRB	LS	TB	RC	HW	HA	CONSENSUS
Mefenamic acid						
PFOS						
2,4,6-Tribromophenol						
Sobetirome						
6-Propyl-2-thiouracil						
Silicristin						
Perchlorate Na						
DiVanadium pentoxide						
TBBPA						
Dibutyl phthalate						
Aspirine						
Pentacholrophenol						
Triclosan						
Ampicilin						
N,N,N',N'-Tetramethyl thiourea (TMTU)						
Ethylene thiourea						
Diclofenac						
Desipramin						
Amiodarone						
Genistein						
Salsalate						
TETRAC						
Ketoconazole						
Niflumic acid						
Sorafenib						
Cd chloride						
2-mercaptobenzothiazole						
Rezorcinol						
Rosmarinic acid						
306	Not solu	Not solu	Not solu	Not solu	Not solu	Not soluble

	solubility) would you recommend a maximum concentration to be tested?
No	
Consensus	The 100 microM is considered sufficient as the highest concentration, provided the chemical is soluble and non cytotoxic.
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	In case two runs are negative, you can stop and conclude on a negative result. If two runs are positive, a third run should be done to conclude. If two runs are discordant, you need to do three extra runs to come to a conclusive result.
Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The group concluded that the assay is valid for the detection of TRalpha and TRbeta agonists.
Question 17	What further work (if considered necessary) would the assessment group recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method or to confirm the acceptance criteria or further development of the data interpretation procedure?)
Consensus	For the agonistic activity, there are other ways (e.g. in silico) to identify chemical structures that are likely to bind the TR pocket. There would be merit to investigate the utility of the assay for the detection of antagonistic activity in addition to agonistic activity. If there was a regulatory need/willingness to develop an OECD TG based on this method, between-lab variability would need to be evaluated. There was some discussion within the assessment group on the relative utility of this assay compared to in silico models.
Question 18	For which chemicals there is sufficient information that they are active or inactive for the mode of action? Please indicate those, so that they can be considered for follow-up validation studies.
Consensus	TETRAC and Sobetirome are the active chemicals and the rest are inactive.

ANNEX 9- CALUX HUMAN THYROID HORMONE RECEPTOR BETA (TRB) REPORTER GENE TRANSACTIVATION MEASURING AGONIST AND ANTAGONIST ACTIVITIES (TR-CALUX)

NON-blinded phase with reference and control chemicals (Part 1)

TR-CALUX Part 1 report: 5 runs are available for few chemicals tested mostly including the reference and control items of the method. Two independent datasets from 2 laboratories were available.

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	Generally, the reproducibility is good across the runs for both agonists and antagonist part of the assay for the BDS dataset (test developer). For the Vitrox data, the agonist part was reproducible across the 5 runs but the data for the antagonist was incomplete. Differences were noted between the two laboratories, questioning the between-laboratories reproducibility of the assay for the antagonistic part, or the suitability of this assay for the antagonist part of the assay.
Question 2	Is the reproducibility across the 5 runs consistent for the chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	Given the limited number of chemicals tested (5 in total including the positive and negative reference chemicals), the assessment group understood this question to evaluate whether this assay works (in terms of its ability to discriminate positives from negatives) or needs further development/optimisation. The response is yes the reproducibility across runs for the chemicals tested is good.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	After discussion within the group and clarification of what the question means and what is the basis for the calculation, the CV is generally below 20% within run (i.e. across replicates), and is thought to be good enough.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	Good for the compounds tested. Assay seems to be workable.
Question 5	Any other observations on the method
Consensus	None
Question 6	Any other observations on the data
Consensus	None

Part 2 report when the chemicals are unblinded:

Part 2 report: 3 runs for Agonism and 2 runs for Antagonism are available for 14 chemicals tested, with 8 concentrations/chemical. Only one laboratory (contrary to Part 1 where two independent data sets from 2 labs were available).

General comment:
On the final classification agonism and antagonism: a feature of the test articles used is that the majority of them were negative in the test (11/14 with 2/14 showing non-specific antagonism). There were only 2/14 positive results. One of the test was inconclusive. Will this eventually impact the assessment of the receiver operating characteristic of the test? Does the small number of positive substances impact the reliability of things like determining the true positive, true positive rate, positive predictive value etc. of the test? While this might not be the question we are trying to answer here, is this worth thinking about for the future>

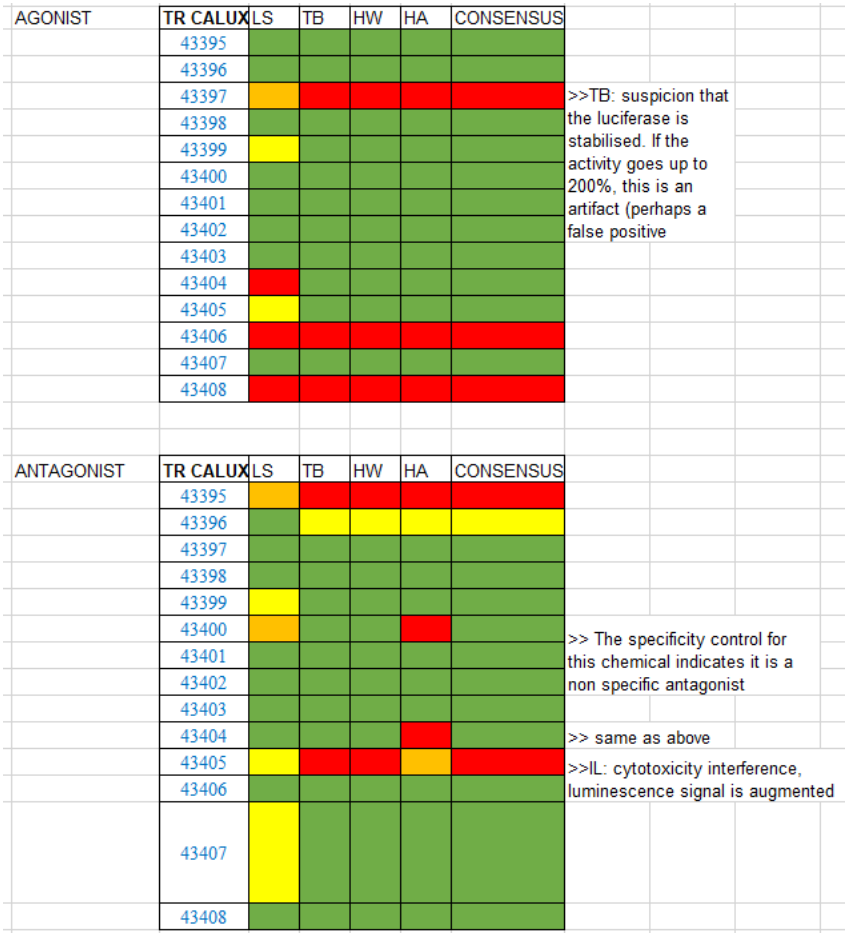
Table 12 Final classification agonism and antagonism

TRβ CALUX (agonism)		anti-TRβ CALUX (antagonism)	
Test item	Classification	Test item	Classification
43395	negative	43395	positive
43396	negative	43396	inconclusive[a]
43397	positive	43397	negative
43398	negative	43398	negative
43399	negative	43399	negative
43400	negative	43400	negative[b]
43401	negative	43401	negative
43402	negative	43402	negative
43403	negative	43403	negative
43404	negative	43404	negative[b]
43405	negative	43405	positive
43406	positive	43406	negative
43407	negative	43407	negative
43408	positive	43408	negative

a – first comprehensive analysis indicated only 1 concentration showing a result below 80% whereas the second comprehensive analysis indicated >2 concentrations causing a relative induction below 80%. Therefore, the compound is classified as inconclusive. Normally, a third analysis has to be performed for confirmation.
b – test item showed an antagonistic response in the anti-TRb CALUX bioassay. Specificity testing showed that the observed antagonistic response of the test item was non-specific and hence, the test item showed non-specific antagonism. Based on specificity test, test item is classified as "negative" based on the antagonistic classifiers.

Part 2 report when the chemicals are blinded

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Agonist Good, Antagonist Fair.
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	Agonist Good, Antagonist Fair (one chemical in particular had inconsistencies between the two runs).
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	For hazard identification, two consecutive concentrations should show a signal with the highest signal above 20% activity for agonist (or 20% decrease in activity for the antagonist) and in the absence of cytotoxicity. Note 1: the SOPs mention 10% activity for the agonist for at least two consecutive data points. Note 2: some chemicals can stabilise luciferase and bias the assay response. This should be considered in the determination of whether a chemical is within or outside the applicability domain of the assay.

	If the assay had some potential use for risk assessment, relevant effective concentrations (e.g. 20%) should be determined from the concentrations tested.
Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	

Part 2 report when the chemicals are unblinded

Question 11	Now that the identity of the chemicals is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	For the agonist part: there was clear consensus that TETRAC and Sobetirome were positive, and almost all other chemicals were negative, but one tested positive: tetrahydroxybenzophenone. The group agreed that it is a false positive result (it is positive based on the criterion proposed by the developer), but this chemical interferes with luciferase in the cytotoxicity assay (stabilisation of luciferase, for which the signal exceeds 150%). For the antagonist part: • PFOS is inconclusive, there are only 2 runs (one of which did not meet the acceptance criteria). An hypothesis was made (in relation to the bell-shaped curve) that PFOS at high concentrations may destroy proteins (luciferase here). This results in a disruption of the signal and the results are inconclusive; • Mefenamic acid is positive; • Triclosan and amiodarone are concluded as negative based on the specificity testing; • Genistein is concluded as positive at high concentrations; • Cd chloride was negative. AGONIST / TR CALUX / LS / TB / HW / HA / CONSENSUS Mefenamic acid PFOS 2,2',4,4'tetrahydroxybenzophenone — FP Dibutyl phthalate Pentachlorophenol Triclosan Ampicilin N,N,N',N'-Tetramethyl thiourea Diclofenac Amiodarone Genistein TETRAC Cd chloride GC1/Sobetirone ANTAGONIST / TR CALUX / LS / TB / HW / HA / CONSENSUS Mefenamic acid PFOS 2,2',4,4'tetrahydroxybenzophenone Dibutyl phthalate Pentachlorophenol Triclosan Ampicilin N,N,N',N'-Tetramethyl thiourea Diclofenac Amiodarone Genistein TETRAC Cd chloride GC1/Sobetirone

Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	Not applicable.
Question 13	How would you judge the specificity of the method?
Consensus	For the agonism, the assay looks specific, (there are interferences with the luciferase signal for both the cytotoxicity and/or stabilisation of the luciferase in the agonism part of the assay. That was the case for tetrahydroxybenzophenone. The antagonism, the picture is less clear (there are interferences with the luciferase signal (destabilisation of luciferase). The DIP criteria would need to be amended to take into account the viability and luciferase activity.
Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	The 100 microM is considered sufficient as the highest concentration, provided the chemical is soluble and non cytotoxic.
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	In case two runs are negative, you can stop and conclude on a negative result. If two runs are positive, a third run should be done to conclude. If two runs are discordant, you need to do three extra runs to come to a conclusive result.
Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The criteria for cell viability measured by luciferase (e.g. below 80% and above 100-150%- precise numbers to be determined by the method developer) for both the agonist and antagonist parts of the assay need to be taken into account in the DIP to determine if a chemical is active or inactive.
Question 17	What further work (if considered necessary) would the assessment group recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method or to confirm the acceptance criteria or further development of the data interpretation procedure?)
Consensus	Further work that is recommended: ○ Test more compounds both for agonism and antagonism. ○ Check the bibliography for compounds that do not have an expected result. Direct effect on the receptor versus indirect effect. ○ Many steps require visual inspection "Visually inspect the microtiter plates, using an inverted microscopy. Check for cloudy wells as an indicator of contamination and verify solubility. Identify samples with cells showing signs of cytotoxicity" or "solubility determination". Dependent on the experimenter. If there was a regulatory need/willingness to develop an OECD TG based on this

	method, between-lab variability would need to be evaluated. There was some discussion within the assessment group on the relative utility of the agonist part of this assay compared to in silico models. For the antagonistic part of the assay, not enough is known to be able to say anything about replacement by in silico models.
Question 18	For which chemicals there is sufficient information that they are active or inactive for the mode of action? Please indicate those, so that they can be considered for follow-up validation studies.
Consensus	For the agonistic part of the assay: TETRAC and Sobetirome are the active chemicals and the rest are inactive. For the antagonistic part of the assay: mefenamic acid can be proposed as active chemical. The other chemicals that showed some antagonist properties (triclosan, amiodarone, PFOS and Genistein) may be un-specific antagonists.

ANNEX 10- MEASUREMENT OF PROLIFERATION OF RAT PITUITARY-DERIVED CELL LINE GH3 (T-SCREEN)

T-SCREEN assay Non-blinded phase with reference and control chemicals (Part 1)

T-SCREEN assay Part 1 report: 6 runs are available for few chemicals tested mostly including the reference and control items of the method.

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	For the agonist part of the assay, the reproducibility across the runs was good. For the antagonist part of the assay, the assessment group felt that the results from Part 1 were ambiguous (as the lab noted in the report) and the assay would need further development in general for the antagonist part of the assay. Although the protocol is not new, it does not seem to have been fully optimised yet for the antagonist part.
Question 2	Is the reproducibility across the 5 runs consistent for the chemicals tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	Reproducibility across runs between the tested chemicals was fair to good.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	The assessment group judged the within-run variability was fair, based on the sub-set of data available. Considering the results in the report, the group judged that the sub-set was probably sufficient to judge variability.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	The assessment group judged the dynamic range to be fair to good (good fr the agonist art and poor for the antagonist part of the assay), considering there were no acceptance criteria pre-defined.
Question 5	Any other observations on the method
Consensus	The group felt that the assay is lacking quality and acceptance criteria for certain test system elements, e.g. cell line characterisation and stability. The assay system generally needs to gain in predictability, stability and performance characterisation.
Question 6	Any other observations on the data
Consensus	Based on the Part 1 data, the assessment group felt that it is not worth moving forward with further validation/chemicals testing of the antagonist part of the assay (not ready). For the agonist part of the assay, the assessment of the group feared that the cell line has drawbacks that make it not very suitable for what it is being used for in this assay (i.e. TR agonism), although Part 2 data and comparison with other assays would be needed to make a judgement.

Part 2 report when the chemicals are blinded

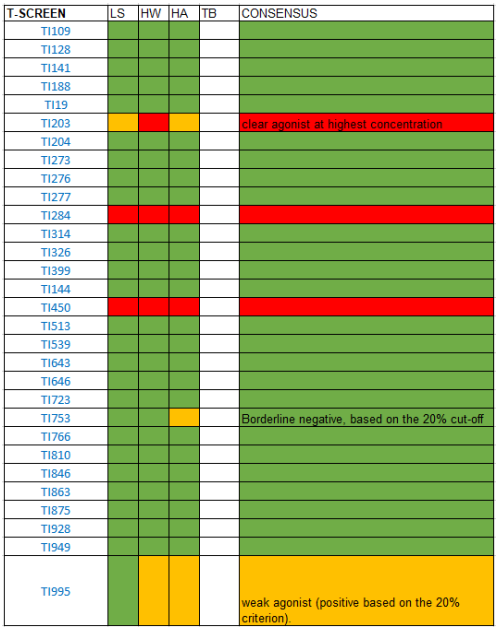
Part 2 report: 3 runs are available for 2 over 30 chemicals tested (1 run for the negative chemicals), with 8 concentrations/chemical.

Reviewer general comment:

- Part 2 report: 30 test items, 2 valid runs, and a 3rd run only for positive items in run 1&2 or for test items with controversial responses in run 1&2.
- SOPs updated for part 2, including procedure for HTS (test acceptance criteria for reference items: see part 1)

Part 2 report when the chemicals are blinded:

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Fair to good.
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	Fair to good.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	For purposes of hazard identification: at non-cytotoxic concentrations, at least 2 consecutive data points with at least ≥20% activity at the highest concentration, with or without a sigmoidal concentration-response curve. For use in risk assessment: A sigmoidal curve allowing the establishment of an appropriate ECx (e.g. EC20 or EC50).

Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	

T-SCREEN	LS	HW	HA	TB	CONSENSUS
TI109					
TI128					
TI141					
TI188					
TI19					
TI203					clear agonist at highest concentration
TI204					
TI273					
TI276					
TI277					
TI284					
TI314					
TI326					
TI399					
TI144					
TI450					
TI513					
TI539					
TI643					
TI646					
TI723					
TI753					Borderline negative, based on the 20% cut-off
TI766					
TI810					
TI846					
TI863					
TI875					
TI928					
TI949					
TI995					weak agonist (positive based on the 20% criterion).

Part 2 report when the chemicals are unblinded:

Question 11	Now that the identity of the chemicals is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	For rosmarinic acid: positive at high concentrations only, which raises some doubt about the specificity of the response (proliferation) in relation to the thyroid mode of action. For other chemicals, there was consensus that previous assessment (blind) is confirmed.

T-SCREEN	LS	HW	HA	TB	CONSENSUS
PFOS					
2-mercaptobenzothiazole					
DiVanadium pentoxide					
TBBPA					
Aspirin					
Rosmarinic Acid					clear agonist at highest concentration
2,2',4,4'-Tetrahydroxybenzophenone					
2,4,6-Tribromophenol					
Tricolsan					
Salsalate					
TETRAC					
N,N,N',N'tetramethylthiourea					
Cd chloride					
Diclofenac					
Perchlorate (sodium)					
Sobetirome					
Resorcinol					
Ketoconazole					
Desipramin					
Pentachlorophenil					
Silicristin					
Genistein					Borderline negative, based on the 20% cut-off
Ampicilin					
Ethylene Thiourea					
6-Propyl-2-thiouracil					
Niflumic acid					
Sorafenib					
Amiodarone					
Mefenamic					
Dibutyl phthalate					weak agonist (positive based on the 20% criterion).

Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	Not applicable.
Question 13	How would you judge the specificity of the method?
Consensus	The assay can respond to different modalities, not just thyroid modalities. The assay is not thyroid receptor-specific, as it measures cell proliferation. The assay is not very specific because growth factors or steroids might stimulate cell proliferation too.

Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	The 100 microM is considered sufficient as the highest concentration, provided the chemical is soluble and non cytotoxic.
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	In case two runs are negative, you can stop and conclude on a negative result. If two runs are positive, a third run should be done to conclude. If two runs are discordant, you need to do three extra runs to come to a conclusive result.
Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The group concluded that the assay is valid for the detection of TR agonists, but the response measured (cell proliferation) is unspecific of thyroid modalities.
Question 17	What further work (if considered necessary) would the assessment group recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method or to confirm the acceptance criteria or further development of the data interpretation procedure?)
Consensus	Other hormones and growth stimulators would need to be tested to assess the specificity of the assay.
Question 18	For which chemicals there is sufficient information that they are active or inactive for the mode of action? Please indicate those, so that they can be considered for follow-up validation studies.
Consensus	TETRAC and sobetirome are clear positive chemicals.

ANNEX 11- MEASUREMENT OF INTRAFOLLICULAR THYROXINE (T4) USING ZEBRAFISH ELEUTHEROEMBRYOS

NON-blinded phase with reference and control chemicals (Part 1)

ZETA Part 1 report: 5 runs are available for 1 chemical tested (KClO4), tested at 3 concentrations, and for the positive control item of the method (Methimazole (MMI).

Question 1	How do you qualify reproducibility across the 5 runs? [Excellent/Good/Fair/Poor]?
Consensus	Generally fair reproducibility across runs given the uncertainty/unclarity in SOPs on the data processing (normalisation). It seems that the number of fish embryos should be increased to have a better perspective on reproducibility; also the way technical replicates are defined needs to be better described in the SOPs.
Question 2	Is the reproducibility across the 5 runs consistent for the chemical tested? In case you see an issue with one chemical, can you identify/hypothesise the reason?
Consensus	Reproducibility across the runs is generally consistent (good) and showing a nice dose-response for the chemical tested.
Question 3	How do you qualify the variability within each run? [Excellent/Good/Fair/Poor]?
Consensus	Fair. Resolution of the data, background staining, use of the microscope and software, placing of the embryos under the microscope, to generally improve the quality of the images could substantially reduce variability within runs. Publications were shared related to background staining issues that are difficult to address: https://link.springer.com/article/10.1007/s10695-018-0488-y This needs to be considered in the SOPs to have acceptance criteria for including a run in the analysis. A higher number of embryos within each run could generally overcome the variability and the ability of the test to detect a statistically significant outcome.
Question 4	How do you qualify the dynamic range (signal to noise ratio) of the method? [Excellent/Good/Fair/Poor]?
Consensus	Fair. There is generally room for improvement, compared to the publications on this assay. The assessment group expressed interest in understanding why such high concentrations were tested, and why the exposure window is so late in the development (embryos have hatched).

Question 5	Any other observations on the method
Consensus	The group discussed the definition of MTC in the scope of this test, based on morphological changes. Some of the listed changes are also indicative of thyroid disruption (e.g. swim bladder inflation). Organisms that show such changes should not be discarded from further evaluation, but such changes should be recorded. Careful considerations are needed for selecting concentrations that are based on acute toxicity testing to ensure sub-lethal concentration are used in the thyroid assay. A table showing the assay workflow, and timing of exposure relative to the duration of the assay, would be nice to have in the SOPs (it is available in the report, section 6.1). The plate layout is not very clear; if they are indeed empty wells, it should be specified. Use of DMSO not needed and far too high level (10 times too high) and is not in accordance with other fish test guidelines in which a maximum of 0.01 % is accepted). This needs to be addressed moving forward. For endocrine assays, efforts should be made to maintain solvent concentration as low as possible to avoid interference. Anaesthesia/euthanasia: it was clarified that the recommended (humane) procedures were used. Data processing and treatment need more prescription in the SOPs.
Question 6	Any other observations on the data
Consensus	Data acceptance criteria and corrective action: what to do with the data when the criteria are not met? Needs clarification in the SOPs. Concern that data resulting from different equipment used or slightly different procedures applied could lead to difficulties in comparing data from different labs. Data normalisation to (solvent) control would help reduce potential discrepancies emanating from slightly different device/equipment/procedures. Such an approach would harmonise the data across laboratories. A power analysis should be performed at some point on these data to find the suitable number of embryos to use in each experimental group and having a sufficient statistical power to detect low decreases in fluorescence.

Part 2 report when the chemicals are blinded:

Part 2 report: 3 runs are available for 12 chemicals tested, with 3 concentrations/chemical.

Question 7	How do you qualify reproducibility across the runs? [Excellent/Good/Fair/Poor]?
Consensus	Within run: probably good. Across runs: fair reproducibility. There is still some concern around the use of the solvent (DMSO concentration of 0.1% which is 10 times higher than what is recommended in aquatic toxicity testing of difficult to test substances).
Question 8	Is the reproducibility across the runs consistent for the chemicals tested? In the case you see an issue with one chemical, please flag it.
Consensus	This question was understood to evaluate whether reproducibility across runs is sufficiently consistent to allow the detection of a chemical effect when there is really a chemical effect expected. The data across runs was generally reproducible when evaluating the test chemicals effect, however, there should be an acceptance criteria for minimum changes between the positive control and the solvent groups.
Question 9	What data interpretation would you apply to determine activity of a chemical tested (i.e. what inhibition rate(s) mean(s) the chemical is active?)
Consensus	Generally, the number of individuals is low and the variability is high, which makes difficult the statistical analysis and interpretation of the data in general. The group also felt that there should be a decision logic how to interpret the data from the three runs, not just the data from individual runs. The group thought that aggregation of data across the 3 runs (each run considered as a technical replicate) would be logical to do to increase the number of animals, but consequently, there would be a single experiment and no other experiment to compare to. There should be biological independent replicates to compare to. The group felt a bit limited in their capacity to propose a clearer data interpretation procedure.
Question 10	For each chemical, can you preliminarily qualify the activity, considering all available information (1) observed response, 2) (in)-soluble concentrations, 3) cytotoxic concentrations and/or concentrations in any way interfering with the test system activity 4) concentrations interfering with the measurement)? Qualify the activity as [weak inhibitor/clear inhibitor/equivocal / negative]
Consensus	Assuming that none of the concentrations tested were toxic to the embryos:

Chemical code	CD	DDP (3 runs)	DDP (pool)	EH	LB	CZ	Consensus	
20								
31							TBC by EH	
343							TBC by EH	
531								
542							TBC by EH	
556							TBC by EH	
703							TBC by EH	
787							TBC by EH	Note: this was the first chemical tested
801							TBC by EH	
838							TBC by EH	
934							TBC by EH	
990							TBC by EH	
		(2 out of 3 rule)			(visual insp.)			

Part 2 report when the chemicals are unblinded:

Question 11	Now that chemicals identity is known, does your initial assessment conform to what you would expect from what is known about the mode of action of the chemicals [weak/clear inhibitor/equivocal/negative]? In case not, can you hypothesise a reason?
Consensus	Salsalate: Provided that this chemical is binding to serum proteins, this chemical should have been active. The concentration tested (around 4 µM) may not have been high enough. To be compared with the in vitro results. Silicristin: Compound difficult to work with, and very quickly metabolised in humans. Was expected to be active because it inhibits MCT8, but not surprised the outcome is equivocal. Genistein: Active as expected. Well known TPO inhibitor with multiple modes of action. Also positive in the DIO1 assay. Resorcinol: TPO inhibitor, metabolised quickly in vivo where the effects are difficult to observe. Active as expected. Perchlorate: NIS inhibitor that is active as expected. LC50 is above 1g/L. Triclosan: TPO inhibitor, active in DIO 1, 2, 3 assays as well. Active as expected. Ampicillin: Result is false positive. The concentration used (300 mg/L) is too high because the LC50 is around 500 mg/L. Systemic toxicity is suspected. 6-PTU: TPO and DIO1 inhibitor, active as expected. The LC50 of PTU is around 630 mg/L, so the concentration used (170 mg/L) is fine and not too high. 2,2',4,4'-Tetrahydroxybenzophenone: TPO inhibitor, expected to be active. Mefenamic acid: Transport inhibitor, active as expected. Aspirin: False positive.Concentration tested (150 mg/L) may be too high. LC50 is around 500 mg/L. 2,4,6-Tribromophenol: TTR binder, active as expected.

	Chemical code	CD	DDP	EH	LB	ZD	Consensus
20	Salsalate						
31	Silicristin						
343	Genistein						
531	Resorcinol						
542	Perchlorate (sodium)						
556	Triclosan						
703	Ampicilin						
787	6-Propyl-2-thiouracil						
801	2,2',4,4'-Tetrahydroxybenzophenone						
838	Mefenamic acid						
934	Aspirine						
990	2,4,6-Tribromophenol						

Question 12	For each of the chemicals that were flagged as 'inconsistent between runs', can you identify/hypothesise the reason?
Consensus	There was great variation between runs for many chemicals. The main hypothesis of the assessment group is that the number of animals is too low and such variations would be less problematic if the number of animals was at least doubled (and this would be feasible, based on the plate layout). A power analysis would help determine the adequate minimum number of animals/run or replicate.
Question 13	How would you judge the specificity of the method?
Consensus	Specificity of the assay is a concern (the two expected negatives ampicillin and aspirin showed activity at the highest concentration). The sensitivity of the assay is good (all expected positives showed activity). More negative chemicals (with no solubility problems) need to be tested at well chosen concentrations to have a better understanding of the specificity of the assay.
Question 14	Looking at the available information (concentration-response curves (shape), solubility) would you recommend a maximum concentration to be tested?
Consensus	The group cannot recommend a maximum tested concentration that fits all chemicals. The maximum tested concentration should be based on a number of considerations including acute toxicity, solubility, and probably harmonised with the other initiatives related to TG 236 updates.
Question 15	Are 3 runs enough/too many given the variability between the runs? How many runs would you recommend the assay to comprise in routine testing?
Consensus	The assessment group agrees that the assay should use at least three runs using more embryos in each run.

Conclusion and recommendation from the assessment group:

Question 16	What is the group conclusion on the validation status of the assay?
Consensus	The conclusion of the assessment group is that the assay is promising but more work is needed. - Number of embryos should be increased; - More negative chemicals should be tested to ascertain the specificity of the assay; - The concentrations tested and solvent concentration should be harmonised with other TG 236-like assays (based on LC10 for example for the MTC) - The spacing of concentrations should be reevaluated and probably be spaced by 1/2Log, or best based on a range finding test to determine toxicity. - Data interpretation procedure for the statistical treatment of data from the 3 runs need further thinking; - Depending on the problem formulation (context of use), this assay might be expanded (testing a fourth concentration,..) for a better dose-response characterisation. - More work is needed on the source of antibodies.
Question 17	What further work (if considered necessary) the assessment group would recommend in order to meet the criteria for adequate validation for the purposes of test guideline development? (e.g. additional data for a certain amount of chemicals to prove the transferability and between lab-reproducibility of the method? or to confirm the acceptance criteria? Further development of the data interpretation procedure?)
Consensus	See above response to Q16.
Question 18	For which chemicals there is sufficient information that they are active or inactive for the mode of action? Please indicate those, so that they can be considered for follow-up validation studies.
Consensus	It is difficult to recommend any negative chemicals at this stage based on the results obtained. For the positive chemicals, the top four chemicals recommended are resorcinol, perchlorate, PTU and mefenamic acid, but all chemicals tested would do a good job as positive chemicals.

Thyroid in vitro methods: assessment reports by the thyroid disruption methods expert group

Reports assessing the validation status of assays from the EU-NETVAL activities

Series on Testing and Assessment

The OECD Test Guidelines Programme established in 2022 a group of experts on methodologies to test and evaluate chemicals for their thyroid disruption properties. Thyroid function is a complex process where multiple biological mechanisms are at stake and represent as many targets of chemical perturbation. Between 2014 and 2024, member countries and dedicated research centres have devoted resources to developing and standardising in vitro assays for thyroid disruption. This document includes consensus reports from OECD on the assessment of the validation status of individual thyroid in vitro assays from the EU-NETVAL activity in the period 2017-2022; assays are listed in chronological order of their progression in the assessment process.