For Official Use **ENV/JM/TG/RD(2015)16**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **03-Mar-2015**

English - Or. English

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

ENV/JM/TG/RD(2015)16
For Official Use

Test Guidelines Programme

COMPILATION OF COMMENTS ON THE FEASIBILITY STUDY FOR MINOR ENHANCEMENTS OF TG 421/422 WITH ED-RELEVANT ENDPOINTS

27th Meeting of the Working Group of National Co-ordinators of the Test Guidelines Programme (WNT)

14-17 April 2015, OECD Headquarters, Paris, France

Nathalie DELRUE
Email: Nathalie.Delrue@oecd.org; Tel +(33) 1 45 24 98 44

JT03371481

Complete document available on OLIS in its original format

English - Or. English

Note from the Secretariat

This document contains the comments and responses to comments received in October 2014 on the draft feasibility report for minor enhancements of TG 421/422 with ED-relevant endpoints. The project for the revision of TG 421/422 was proposed by Denmark in 2013 and the draft feasibility report submitted to the WNT for comments in September 2014. The revised draft feasibility report is available in document [ENV/JM/TG(2015)24].

ACTION REQUIRED ***The Working Group of the National Coordinators of the Test Guidelines Programme is invited to take note of responses to comments.***

Comments on the draft feasibility study for the update of TG 421 and 422 with ED related endpoints

Comments were received in October 2014 from Denmark, Italy, Japan, Netherlands, Spain, Sweden, Switzerland, UK, US and BIAC

Member country	General Comments	Responses
Italy	Both the TGs 421/422 should be used in specific situations where high exposure and high volume of production are expected but comparably no toxicological data are available on the effects of test chemicals. In other situations, a screening battery of in vitro tests should be preferable. Since TGs have to screen chemicals for the reproductive/developmental toxicity, they should include specific evaluation of the potential maternal toxicity. The relevance and specificity of any effects in F1 should be weighted on the basis of maternal data. In particular, for TG 421 histopathological examination of the main target organs - e.g. liver and kidney in dams – for maternal toxicity is suggested. The ability of TG in screening endocrine disrupting effects should be increased by considering the histopathological analysis of both maternal and F1 thyroid to strengthen the serum biomarker evaluation.	At least general clinical observations, body weight and food consumption of the maternal animals are required (more in TG 422). Thyroid histopathological examinations in dams and pups have been reflected in the new version of the TGs and in the review.
Japan	Examinations of females with or without pregnancy seem to be generally inadequate compared to males. Blood (serum or plasma) hormone measurements (T3/T4 and/or TSH) are considered to be informative in a case that any endpoint indicating thyroid hormone-related changes such as growth retardation at doses of no maternal toxic effects is observed. If no effects indicating the thyroid effect are observed, hormone measurement is not necessary. Based on effects of a test substance on the endocrine system, appropriate and flexible approaches of hormone analysis should be addressed.	*Clarification needed?* Thyroid hormone measurement based on triggers – issue of the sensitivity of growth retardation in pups as a trigger. Thank you for this comment. However, it seems that growth retardation as internal trigger to T4 measurements is not a sensitive trigger. Papers on studies of PTU, OMC and triclosan show significant and marked decreases in T4 levels at PND (postnatal day) 16 offspring at doses below those that

Member country	General Comments	Responses
		induce reduced birth weight (see references below). Moreover, PTU in one study did not result in growth retardation at all. There are also clear behavioural effects of PTU without any effect on birth weight. Moreover, a study on PFOS by Yu et al, shows decreased T4 and no effects on juvenile weight. **Ref**. Axelstad M, Hansen PR, Boberg J, Bonnichsen M, Nellemann C, Lund SP, Hougaard KS, Hass U (2008) Developmental neurotoxicity of Propylthiouracil (PTU) in rats: Relationship between transient hypothyroxinemia during development and long-lasting behavioural and functional changes. Toxicol Applied Pharmacol 232, 1-13 Axelstad M, Boberg J, Hougaard KS, Christiansen S, Jacobsen PR, Mandrup KR, Nellemann C, Lund SP, Hass U. (2011) Effects of pre- and postnatal exposure to the UV-filter octyl methoxycinnamate (OMC) on the reproductive, auditory and neurological development of rat offspring. Toxicol Appl Pharmacol. 250, 278-90. Axelstad M, Boberg J, Vinggaard AM, Christiansen S, Hass U.(2013). Triclosan exposure reduces thyroxine levels in pregnant and lactating rat dams and in directly exposed offspring. Food Chem Toxicol.59, 534-540.

Member country	General Comments	Responses
	Females in a non-pregnant group should be checked their vaginal smear throughout study period. In addition, histopathological confirmation in the ovary, uterus or vagina is useful (Please see comment on Para.54 in TG 422 compilation of comments).	Yu WG1, Liu W, Jin YH, Liu XH, Wang FQ, Liu L, Nakayama SF (2009) Prenatal and postnatal impact of perfluorooctane sulfonate (PFOS) on rat development: a cross-foster study on chemical burden and thyroid hormone system. Environ Sci Technol. 2009 Nov 1;43(21):8416-22. doi: 10.1021/es901602d. Modulation to estrous cyclicity proposed to be investigated via examination of vaginal smears. See revised paragraph in TG 422.
Netherlands	We support the update of the TGs with the indicated parameters.	Comment noted.
Spain	Based on the outcome of the feasibility report, the updated versions of TG 421 and TG 422 have our approval. We support submission for approval at the next WNT meeting.	Comment noted.
Sweden	Feasibility study for minor enhancements of TG 421/422 – the document gives a good overview and description of the suggested endpoints and how and why they should be added to the test guidelines. I support the suggested improvements of the TG 421 and TG422	Comment noted.
Switzerland	We agree to engage in the next phase of the project, i.e. the update of TG 421 and TG 422, implying a potential submission for approval of the draft updated TGs at the next WNT meeting. This is a well-presented proposal, which would increase the usefulness of TGs 421 and 422 at little additional cost. **National position:** Agree	Comment noted.
UK	● The UK supports the revisions to TG 421 and 422 and the detailed amendments and agrees	The introductions of TG 421/422

Member country	General Comments	Responses
	that this should be moved forward. However all the recent discussions and background to the proposed revisions have not been reflected in the introduction. **National position:** Agree, new text required	were revised to better reflect the background behind the current update.
	● I support the addition of the new endpoints to these TGs, they are valuable additions informing on potential endocrine activity whilst not adding substantially to the complexity of the studies. It seems to me that there may have been an opportunity to include further endpoints of thyroid hormone disruption in both TG 421 and 422 (thyroid weight and histopathology in dams perhaps, 422 does include thyroid histopathology), I wonder if this was considered? **National position:** Please can the expert group discuss?	See revised TGs as well as the review for clarification after thyroid discussions in EG. Not in line with the previous comment – need for national position.
	● Response to Proposal to add Analyses for Thyroxine (T4) and Thyroid Stimulating Hormone (TSH) as Mandatory Endpoints to OECD Test Guidelines 421 and 422: See annex 1 of this compilation of comments **National position:** Please can the expert group discuss.	See revised TGs as well as the review for clarification after thyroid discussions in EG.
US	My comments pertain to the report entitled, "Feasibility study for minor enhancements of TG 421/422 (Reproduction/Developmental Toxicity Screening Test)/Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test) with ED-relevant endpoints." Dated 28 August 2014. Inclusion of anogenital distance, nipple retention, and thyroid hormone levels into the OECD TG 421 and 422 are justified and sufficiently validated across different laboratories and paradigms. Their relevance to human health effects has been recognized by the general reproductive and	Comment noted.

Member country	General Comments	Responses
	developmental toxicology community. I support their incorporation into OECD TG 421. However, I do not think the same degree of support, justification, and validation exists for including an external exam for male genital malformation early after birth (Day 0 to Day 6 postnatal) as described in the report. The data used to support this approach (starting paragraph 58 of the report) are composed of pilot data from a single laboratory. Furthermore, there is apparently no clear consensus regarding the statistical methods of this approach. There is already an examination of the external genitalia of pups at animal termination included in TG 421 and 422. At this time, it is unclear what added sensitivity and value is obtained by an examination in the early postnatal period to detect external malformations. Male malformations may be difficult to detect at this young stage due to the underdeveloped nature of rodent genitalia at birth. TG 443 (EOGRTS) measures the presence of male malformations at sexual maturation partly due to the difficulty in detecting some malformations until the animals have reached a more mature phenotype. Until inter-lab validation has been performed, it is premature to incorporate a very early postnatal measurement of external malformations. It is quite possible for malformations to be underreported due to the difficult of detection which may affect the interpretation of the study data. **National position:** Expert group should consider whether it is appropriate to include examination of males on the basis of data from one lab.	As indicated in the feasibility report, this is already to be done as soon as possible after delivery (new para 39 TG 421) and at termination (new para 47 TG 421). In para 475 in TG 421 is only added: "Particular attention should be paid to the external reproductive genitals which should be examined for signs of altered development". This text proposed to be added in the revised TG 421 and 422 in relation to abnormalities is modified from para 30 in OECD TG 414. The provided addition is for clarification and not for classification of malformations vs. variations which is applicable to 414 but not 421 or 422.
BIAC	The feasibility study only consider if the endpoints can be added to the screening assays. It does not address how the proposed changes will affect the current endpoints for repeated exposure toxicity and/or reproductive/developmental toxicity. For example, factors such as not standardizing litter sizes may increase the variability in pup body weights and litter sizes, making these types of endpoints less sensitive. There should be a section in the document that addresses the potential impact on all of the currently required endpoints in the TG 421/422 studies. This importance not to affect the validity of the current endpoints being collected.	It is well known that litter size may have influence on pup body weight. This can be addressed in the statistical analyses by including litter size as a covariate in the analyses of pup body weight. This will be added in the relevant sections on statistical analysis (new para 51 in TG 421 and 72 in TG 422).
BIAC	Thyroid hormone samples should NOT be collected on PND 13. There are age-related changes	See revised TGs as well as the

Member country	General Comments	Responses
	in thyroid hormone levels in rat pups during lactation (i.e., T4 levels are very low on PND 4, increase ~13X by PND 15, then decrease ~4X by PND 21 to levels similar to adult animals. Per Zoeller and Tan, Crit. Rev. Toxicol. 37:195-210, 2007: T4 levels in normal rat pups are in the range of 0.5 to 1.0 μg/dl on PND 4 (Goldey et al., 1995; Zoeller et al., 2000), rising to 8 to 12 μg/dl on PND 15, then declining to adult levels of approximately 3 μg/dl by PND 21). Developmental delays (possibly decreased pup body weights and/or differences in litter size) could affect T4 levels by altering the transition to adult levels.Decreases in nutritional status/food consumption and/or stress can alter thyroid hormone levels (e.g., Eales, Amer. Zool. 28:351-362, 1988; Döhler et al., Pharmac. Ther. 5:305-318, 1979). Lactation induces a slightly hypothyroid state in rats and LD 13 is in the phase of maximum milk production; thus, thyroid status is already below baseline levels (Quevedo-Corona et al., Life Sci. 66:2013-2021, 2000). PND 13 is too dynamic for thyroid hormone assessment, which will both increase the likelihood of non-specific responses and increase the variability in hromone measurements. If included, thyroid assessments should be done earlier (PND 4) or later (PND 21) when thyroid hormone levels are more stable. Furthermore, changes in thyroid hormones are often adaptive; therefore, changes should be interpreted with consideration of thyroid weight and histopathology.	review for clarification after thyroid discussions in EG. Extracted from the comment: "If included, thyroid assessments should be done earlier (PND 4) or later (PND 21) when thyroid hormone levels are more stable." But samples are taken at termination and termination is at PND 13.
BIAC	With such small sample sizes, thyroid hormone levels have the potential to generate many positive results, because effects are not specific to the thyroid, resulting in needless follow-up studies to evaluate potential thyroid effects.	See revised TGs as well as the review for clarification after thyroid discussions in EG.
BIAC	Thyroid hormone data should be interpreted in conjunction with thyroid weight and histopathology. Thyroid hormone can be modulated by a number of non-specific variables (e.g., stress) and changes in hormone levels may be adaptive; thus, additional data is needed to interpret the biological significance of any hormone measurement changes.	See above
BIAC	Nipple/areolae counts are somewhat subjective as the observer is, in some cases, counting subtle differences in shading of the skin. This contributes to inter-study variability as well as inter-laboratory variability (e.g., some laboratories have higher background nipple counts in males than other labs). Thus, differences in nipple/areolae counts should be interpreted with other endpoints that are potentially indicative of antiandrogenic effects; subtle differences in nipple/areolae counts in the absence of supporting data may not be biologically meaningful.	Reference is made to GD 151 that indicates that the presence of nipples should be measured when they are obvious, thus limiting the subjectivity. In addition, the proposal is to extend the study duration to enable optimal observation time period i.e. PND

Member country	General Comments	Responses
	Given the limited sample sizes in this study, NOAEL should not be based on this parameter alone; perhaps follow-up studies would be warranted.	13. This type of discussion goes beyond the scope of the TG. NOAEL is set by expert judgement after consideration and analysis of study results.

Member country	Specific comments	Responses
	Paragraph 10	
Denmark	Text should be revised to (see in bold): 10. Assessment of AGD and NR are mandatory in TG 443 and could probably easily be included in the TG 421/422. For the examination of NR it **is** ~~may~~, however, ~~be~~ needed to extend the study period in 421/422 from PND 4 to PND 12 or13 to examine this endpoint at the optimal time period. **National position:** Please revise	Done
BIAC	Agreed - Prolonging the study would be needed to incorporate nipple retention.	Noted
	Paragraph 11	
Denmark	Text should be revised to (see in bold): 11. The OECD TG 407 (Repeated dose 28- day oral toxicity study in rodents) has been updated in 2008. The assay has been validated for some endocrine endpoints but the sensitivity of the assay is not sufficient to identify all EATS-mediated EDs. The validation of the assay (OECD,	Done

	2006) showed that it identified strong and moderate EDs acting through the ER and AR; and EDs weakly and strongly affecting thyroid function. It was relatively insensitive to weak EDs acting through the ER and AR. This assay also **has ~~ve~~** some optional endpoints such as uterine and ovary weight, Changes in vaginal smears, histopathologic changes in mammary gland histopathology as well as serum T3, T4, TSH as well as thyroid weight. **Comment:** It would be nice with some sort of conclusion to this sentence, like." …which can be used if there is additional concern…" or " "…which are however rarely assessed.." or some other conclusion. Without this, this paragraph in my opinion seems a little unfinished. **National position:** Please revise	Text is revised to: …well as serum T3, T4, TSH as well as thyroid weight **which can be examined if there is additional concern**.
	Paragraph 12	
Denmark	Text changes is in bold: The new extended one-generation reproductive toxicity study (EOGRTS)…. as it also uses reduced animal numbers **if conducted without F2**, it is expected that it will often replace OECD TG 416 …for mammalian re-productive toxicity testing (GD 150). … Effects on the developing nervous and immune systems are also assessed **by the DNT and DIT cohorts**. **National position:** Please consider adding the text in bold	Done
BIAC	A two gen study protocol could also be modified to include ED endpoints in the event a compound does not have neuro or immune system concerns and an EOGRTS design is not warranted.	Such design would not be covered by MAD. No change was made to the text.
	Paragraph 13	
BIAC	Would evaluation of genital malformations at time of sexual maturity be preferable in terms of identification of the external abnormalities or distinguishing ED-related findings from developmental delay?	Correct that male malformations may be easier to detect in sexually mature animals due to the underdeveloped nature of rodent genitalia at birth. However, it is

		possible to detect genital abnormalities in offspring around day 13 and the results of this predicts the likelihood for malformations in mature animals quite well (Christiansen et al 2008). Ref. Christiansen S, Scholze M, Axelstad M, Boberg J, Kortenkamp A, Hass U. Combined exposure to anti-androgens causes markedly increased frequencies of hypospadias in the rat. International Journal of Andrology 2008;31:241–8.
	Paragraph 17	
Denmark	…Christiansen et al 2008 – please add more refs. **National position:** Please add more references here.	The following references are provided (Bowman et al., 2003, McIntyre et al., 2002; Welsh et al 2008).
	Paragraph 26-28	
BIAC	AGD serves as a biomarker of androgenicity/antiandrogenicity in male rats; however, the human relevance of this endpoint is highly speculative. For example, finasteride is capable of producing marked effects on AGD in rats, but not primates. In an excerpt from Clark (An Evaluation and Interpretation of Reproductive Endpoints for Human Health Risk Assessment, ILSI press, 1999): "...there are no reports of decreased anogenital distance in Type 2 pseudohermaphrocitic men who lack 5alpha-reductase or in rhesus monkeys treated with finasteride. The reason for these differences between rodents and primates may be because development of the external genitalia in humans (Scholly et al., 1980; Glenister, 1958; Spaulding, 1921) differs from that in rodents (Anderson and Clark, 1990). In rats, the male and female external genitalia are distinct on day 19 of gestation based on anogenital distance but the urethral plate is located on the ventral side of the genital tubercle in both sexes. By day 21, the urethral plate, still contained within the genital tubercle, has opened to form the urethral groove in females but not in males. The region measured by anogenital distance excludes the urethral groove in females and the urethral plate in	Thank you for this input; we made some revisions in para. 27 and 28

	males. In contrast, in human embryos, the urethral plate has opened to form the urethral groove in both males and females in the indifferent stage (e.g., gestation week 7) and it is located between the anus and the genital tubercle. By gestational week 12, the urogenital folds have fused in males to form the penile urethra whereas the urethral groove persists in females. Thus, a different sequence of events is occurring in the region between the anus and the genital tubercle, which is measured in rodents as anogenital distance."	
	Paragraph 27	
BIAC	Swan et al. studies describe preliminary findings requiring follow-up based on the authors' own descriptions. These publications should not be cited as definitive evidence of effects of phthalates in humans. Furthermore, when the National Toxicology Program Center for the Evaluation of Risks to Human Reproduction (CERHR) evaluated phthalates, there were several questions about Swan's work that remain unanswered. These data should not be used as evidence of effects in humans.	New para 27 does not refer to the preliminary findings in the Swan studies (see comment above)
	Paragraph 33	
Denmark	This paragraph should in my opinion also include an introduction of 'areola' and an explanation of the difference between nipples and areola. **National position:** Please add such introduction	Have been addressed in para. 36
Switzerland	Change "DHT" to "dihydrotestosterone (DHT)" **National position:** Agree	Done
	Paragraph 34	
BIAC	Nipple Retention: Considering the 10 day extension to assess nipple retention was a cost: benefit ratio conducted to determine if this information was worth the additional cost?	In the conclusion it is stated that: However, in almost 30% of the studies (6 out of 21) is nipple retention more sensitive than AGD. Therefore, inclusion of both AGD

		and nipple retention will provide an increased ability for evaluating the potential endocrine disrupting activity of a substance compared to having only data for AGD. This is especially relevant in cases where equivocal AGD data are found.
	Paragraph 40	
Denmark	The data are from 20 Copenhagen studies please consider deleting Copenhagen: The data are from 20 ~~Copenhagen~~ studies **National position:** Please consider deleting Copenhagen if possible also in para 25. Please also delete in appendix 1 If not deleted please explain what Copenhagen and non-Copenhagen means.	In Appendix 1. Power Simulations of… Copenhagen studies and non-Copenhagen studies is used. The term is explained in para 25 as: The results based on both the Copenhagen studies (data from Division of Toxicology and Risk Assessment, National Food Institute, Technical University of Denmark) and the non-Copenhagen studies (Data from other labs) shows that… And para 40: The data are from 20 Copenhagen studies (see para 25).
NL	It would be valuable to add the background level for NR for the 'standard' strains (e.g. Wistar, Wistar Han, Sprague Dawley). **National position**: Agree	This has been addressed in para. 40.
	Paragraph 43	
BIAC	Utilizing the NOAEL from Nipple Retention as a point of departure, in the absence of other supporting endpoints is somewhat concerning. This endpoint, while sensitive for anti-androgens, should be used in conjunction with other complementary and redundant endpoints within a study	See also response to last general comment from BIAC. The wording "should be considered in setting a NOAEL" allows flexibility

	to confirm a potential disruption in androgen signaling.	
	Paragraph 45	
Denmark	The argumentation from the discussion/conclusion on including NR, even though it means that the study has to be prolonged, should in my opinion also be mentioned here. **National position:** Please include this argumentation	The end of para 45 is changed to: For the OECD TGs 421/422 an extension of the testing period from postnatal day 4 to 12 or 13, i.e. 9-10 day is necessary as nipple retention has to be assessed on postnatal day 12 or 13.
	Paragraph 46	
Denmark	A quantitative count in male pups is required as a qualitative assessment only (presence/absence) of nipples/areolae may be rather insensitive. Could be changed to: A quantitative count in male pups is required as a qualitative assessment only (presence/absence) of nipples/areolae **is regarded being to** insensitive. **National position:** Please add text in bold	Done
	Paragraph 51	
Sweden	Paragraph 51 describes the common practice of pooling blood samples from pups – the suggestion, that I fully support, in the TG is to pool pups blood per sex and litter (Appendix 2a paragraph 35a and Appendix 2b paragraph 44a in TG 422). I think that it would be good (not necessary) if some general information about sex differences in thyroid hormone levels are given, maybe in the human relevance section.	Generally sex differences are not seen in pups before sexual maturation. No change in text.
	Paragraph 52	
BIAC	Minimal effect differences that can be detected for thyroid hormone (i.e., effect sizes) were	See revised TGs as well as the review for clarification after thyroid

	based on a Coefficient of Variation of 20%, which was reported for adult animals from PND 41-53. Given that lactation is a very dynamic period for thyroid hormone levels (see above), it is highly likely that the Coefficient of Variation will be markedly higher in PND 13 pups. To accurately evaluate the effect size for these data, thyroid hormone levels from PND 13 pups must be utilized. The power calculations presented in the feasibility document are unsupported and should not be used to justify inclusion of this endpoint in OECD 421/422 studies.	discussions in EG. A new power analysis of T4 and CVs have been included (appendix 1b).
	Paragraph 53	
Denmark	Please delete the sentence during recent years… This sentence states the same as the previous one and should therefore be deleted **National position:** Please delete	Done
	Paragraph 54	
Denmark	Please delete: induced by various thyroidal or extra-thyroidal initiating events I would recommend deleting this last part of the sentence, as I think it is very difficult for people to understand exactly what it means. **National position:** Please delete Please end the last sentence with: making rat data on this endpoint relevant for human health risk assessment. **National position:** Please add the sentence	Done Done
	Paragraph 55	
Denmark	Please revise to:	

	If blood samples for assessment of thyroid hormones in dams and pups are taken at termination of the study (i.e. PND 4 or PND 13-14 if assessment of NR is included) this leads to no concern for animal welfare, as trunk blood can be collected at the time of sacrifice. In adult animals in TG407, fasted blood samples are to be used, and fasting for 20-24 hours in adults only leads to minor concern for animal welfare. However, these blood samples are proposed only to be taken in TG 421/422 if they have not already been taken in a TG 407 study investigating the same compound. **National position:** Revise And clarify what is meant by: These studies include relatively similar number of adult animals (5-8 per dose per sex) and therefore, the overall animal welfare considerations will not increase by this assessment and are evaluated as minor **National position:** Please clarify	Done Clarified. The text have been revised to: The TGs 421/422 studies include relatively similar number of adult animals as TG 407 (5-8 per dose per sex) and therefore… Done
NL	For OECD TG 421, it would increase discomfort for the parental animals as normally, these would not be fasted overnight before necropsy. In addition, many labs take blood samples before the animals go into necropsy (so not as part of the necropsy procedure), and this would lead to additional discomfort as the animals would awake from anaesthesia. **National position:** Fasting is aspect that is included in several OECD TGs. The information obtained in these parameters outweighs the discomfort.	Please see the clarification in revised paragraphs in relation to fasting.
	Paragraph 56	
Denmark	Please revise to: There are standardized OECD test methods for assessing thyroid hormones and the performed sensitivity analysis indicates that data on TH level changes of about 25 % and more, will be able	Done.

	to be obtained with the number of litters per group in the TGs 421/422. This may at first glance not seem very sensitive. However due to the previously explained species differences in the thyroid hormone system between rats and human, exposure to thyroid disrupting compounds will probably in many cases lead to a greater response on hormone levels in rats than would be expected in humans. Therefore inclusion of TH measurements in the TG421/422 is not as insensitive as it would seem at first glance and is still conserved relevant. Due to the adverse effects seen in humans after developmental hypothyroidism, this endpoint is of high human relevance and there are no concerns for animal welfare related to the assessment of this endpoint as long as blood sampling is done in animals that are being sacrificed anyway. This all supports that assessment of thyroid hormones can be included in TGs 421/422. **National position:** Please Revise	
BIAC	Statistical power regardless, limiting thyroid samples to 8/group leads to a high potential for false positives, especially if no attention is paid to the pattern of findings and supporting evidence based on thyroid weights and histopathology. A statistically significant change in a single thyroid hormone should not be regarded as evidence of an adverse effect particularly when sample size is limited, due to the high variability in these parameters (in both adults and pups, in our experience). In fact, thyroid hormone findings unsupported by histopathological changes should not be characterized as adverse, and hypertrophy and colloid depletion are adaptive rather than adverse.	See revised TGs as well as the review for clarification after thyroid discussions in EG. A new power analysis of T4 and CVs has been included (appendix 1b).
	Paragraph 60&61	
US	The validation of detecting male sexual organ external malformations appears to have been conducted by a single lab. Furthermore, the data are reported to be "unpublished," which implies the data have not been peer-reviewed. There is no indication of whether other labs besides this single lab would be able to adequately perform the external exam as described. Inter-lab validation should be performed. The authors explicitly acknowledge that the statistical analyses for external malformations were performed incorrectly. Furthermore, the authors state that a method to incorporate litter effects has not been designed or validated. Until appropriate statistical methods are identified for this endpoint, it is premature to incorporate this endpoint into the TG 421/422.	See above, response to US general comments.

	National position: Expert group should consider whether it is appropriate to include examination of males on the basis of data from one lab.	
	Paragraph 61	
BIAC	The analyses presented in Pt 61 are useless unless analyzed on a per litter basis. If there is no statistical method for doing this, one could refer to developmental toxicity studies where an n of 20 is considered appropriate for identifying exposure related abnormalities. The standard practice for evaluating developmental abnormalities suggests that if one did record increased genital abnormalities in these screening studies, it should at most be regarded as something to evaluate in more extensive studies and not be used to define regulatory NOAELs.	We find that the analyses illustrate the sensitivity of for detecting abnormalities.
	Paragraph 71	
Denmark	If all four of these endpoints are included in TG 421/422, the animal welfare considerations are evaluated as minor. Could be changed to: **Inclusion of** all four ~~of these~~ endpoints ~~are included~~ in TG 421/422 **does not trigger any**~~, the~~ animal welfare **concerns.** ~~considerations are evaluated as minor.~~ **National position:** Please consider changing to the bold text.	Done
	Paragraph 74	
Denmark	In conclusion, it is feasible to make these minor enhancements of TG 421/422… Please consider revising to: In conclusion, it is feasible to make the **proposed**~~se~~ minor enhancements of TG 421/422 … **National position:** Please change the wording	Done
	Appendix 1: Tables 1 and 2	

Switzerland	Please clarify which units we are using to describe Anogenital Distance (AGD) in figure 1 and tables 1 and 2: Figure 1: Relationship between birth weight and AGD in male controls. Table 1: AGD – Dose-Response summary (Copenhagen studies): Table 2: AGD – Dose-Response summary (Non-Copenhagen studies) In table 1 AGD-mean values range from 11-14 units in females and 20-25 units in males. In figure 1 AGD-mean values in male controls at birth show similar numeric results. However, the non-Copenhagen AGD data (Table 2) indicate much lower numeric values. Means are about 4 for males and 2 for females. They are presumably in mm. **National position:** Point for clarification	The AGD in the Copenhagen study is given as units (could be converted to mm) and in the non-Copenhagen studies the AGD are given in mm. This has been revised in appendix 1.
	Appendix 3	
BIAC	"Kontrol" in the graph on p.74 should be "control".	Done

Annex 1 – additional comments from UK
(Provided by individual members from a UK industrial reproductive toxicology group)

Response from lead country (LC): These comments have been considered together with the UK national position and by the EG. In addition, a few of the comments have been addressed by LC in the text below (highlighted in grey)

Response to Proposal to add Analyses for Thyroxine (T_4) and Thyroid Stimulating Hormone (TSH) as Mandatory Endpoints to OECD Test Guidelines 421 and 422

It has been proposed that the following paragraph is added to both Test Guidelines (Paragraphs 35a and 44a in Test Guidelines 421 and 422, respectively):

"Blood samples from a defined site are taken from all pups and all dams at pup PND 13, at termination, stored under appropriate conditions and assessed for serum levels for thyroid hormones (T4 and TSH). Pup blood can be pooled by sex per litter for thyroid hormone analyses."

Listed below are reasons why this proposal is not appropriate.

1. Lack of Statistical Power

Both guidelines specify starting the study with 10 females per dose level, with the expectation that this will yield at least 8 pregnant females, and therefore yield 8 litters. The review document Feasibility Study for Minor Enhancements of TG 421/422 with Endocrine Disrupter-relevant Endpoints comments that no detailed statistical power calculations have been performed with respect to measurement of T_4 or TSH, but have presented some information on estimated statistical power, making a number of assumptions regarding the coefficient of variation and with 3 different group sizes for dams and litters per dose level (the below table is copied from the review document):

Minimal statistical detection limit* for three different data scenarios (CV=10%, 15% and 20%) - Reductions from the control mean that can be detected at given litter sizes (N=8, 12 and 20) and error rates α=5% and β=20% (i.e. 80% power).

Coefficient of Variation (CV)	Litter Size		
	N =8	N =10	N =12
10%	13%	10%	8%
15%	20%	16%	12%
20%	26%	21%	16%

*t-test, one-sided, balanced litter design.

Based on this analysis, the authors state:

"Generally, for small or moderate sample sizes the chance of detecting small hormonal changes is rather low, even if low data variability is expected. At least for adult rats it is more likely to observe CVs of around 20%, and in these cases only high sample sizes would ensure the detection of at least 15% hormonal changes"

We agree with the notion that these studies have very low statistical power for detection of effects on T_4 and TSH, but disagree with the analysis as presented for a number of reasons:

a. The only applicable group size (N) is 8. The numbers presented for N=10 and N=12 should not be considered as increasing the group size is contrary to animal welfare concerns as detailed in the feasibility study.

Higher coefficients of variation should be considered. The table presents scenarios where CV = 10, 15 or 20%. Syngenta believes that CV levels of 10 and 15% are unrealistically low and that these scenarios should not be considered. This position is supported by the data presented for young adult Sprague Dawley rats in OPPTS Test Guidelines 890.1450: Pubertal Development and Thyroid Function in Intact Juvenile/Peripubertal Female Rats and 890.1500: Pubertal Development and Thyroid Function in Intact Juvenile/Peripubertal Male Rats:

Parameter	Sex	Mean	Standard Deviation	Coefficient of Variation
T_4	Female	4.03 µg/dL	0.67	21.38
T_4	Male	5.716 µg/dL	0.83	18.27
TSH	Male	14.162 ng/mL	4.975	34.04

Based on these data (which reflect the data from a number of studies), we believe that the lowest coefficient of variation to be used for more realistic scenarios for power calculation would be 20%.

We agree that a two-sided ANOVA analysis should be done, since hormone levels the principle can be both higher and lower and t-tests are not optimal for comparisons with several doses. The argument that the CV of 20 (rather than 10 and 15) is the most relevant is supported by the table above. Revisions will be done to reflect this.

b. The analysis as presented uses a one-sided statistical test (t-test). Typically, for studies of this design where the test substance is being 'screened' for an effect with no assumed or *a priori* knowledge of expected effects, a two-sided test it more appropriate as this considers that exposure to the test substance could both increase or decrease the levels of T_4 and/or TSH. Using a one-sided test in this analysis overestimates the minimum statistical detection limit. In addition, the t-test is not an appropriate test for a study with multiple treatment groups (as recommended in both OECD TG 421 and 422) as it does not correct for multiple pairwise comparisons (i.e. each treatment group being compared to the control). Not correcting for multiple comparisons overestimates the minimum statistical detection limit. Syngenta notes that for the power calculations for anogenital distance (AGD) in the review document, two-sided tests were used and consideration was given to correction for multiple pairwise comparisons.

Based on the points made above, the addition of analysis for T_4 and TSH into OECD TGs 421 and 422 is not supportable and the study designs do not have sufficient statistical power to detect changes in these parameters.

There are standardized OECD test methods for assessing thyroid hormones and the performed sensitivity analysis indicates that data on TH level changes of about 25 % and more, will be able to be obtained with the number of litters per group in the TGs 421/422. This may at first glance not seem very sensitive. However due to the previously explained species differences in the thyroid hormone system between rats and human, exposure to thyroid disrupting compounds may lead to a greater response on hormone levels in rats than would be expected in humans. Therefore inclusion of TH measurements in the TG421/422 is not as insensitive as it would seem at first glance and is still conserved relevant. Due to the adverse effects seen in humans after developmental hypothyroidism, this endpoint is of high human relevance and there are no concerns for animal welfare related to the assessment of this endpoint as long as blood sampling is done in animals that are being sacrificed anyway.

2. Study Design, Technical and Logistical Concerns

It is well known that, like most hormones, T_4 and TSH levels show considerable variation between individual animals as well as considerable variation throughout the day owing to normal circadian

cycling (e.g. Ottenweller and Hedge, 1982). This high level of variation introduces a number of study design and logistical challenges that render incorporation of their measurement into OECD TGs 421 and 422 inappropriate.

a. The first challenge is minimising the impact of the normal diurnal variation by ensuring that the animals are terminated and bled in the narrowest time window possible. In OPPTS 890.1450 and OPPTS 890.1450, it recommends that in order to control for this diurnal variation, all animals on study are terminated within a time window as narrow as 2 h (latest time for final dose administrations: 0900; latest time for termination beginning: 1100 (2 h after dosing); all termination to be completed by: 1300). OPPTS 890.1450 and OPPTS 890.1500 typically require 45 animals in total. With the current proposals for OECD TGs 421 and 422, a considerably greater number of animals (~480, based on control + 3 treated groups, with 8 dams per group and an average litter size of 14 for Sprague Dawley rats (http://www.criver.com/files/pdfs/rms/cd/rm_rm_r_tox_studies_crlcd_br_rat.aspx) would have to be terminated and bled for T_4 and TSH analyses and therefore controlling for diurnal variation would be extremely difficult, if not impossible.
Minimizing the impact of diurnal variation when measuring thyroid hormone levels can be done by distributing the sacrificed animals from different dose groups over the entire sacrifice period. Sacrifice of additional animals means that they cannot be terminated within a 2 hour time window, but even though increasing this time window increases the variation in data, controlling for diurnal variation is not impossible as long as timing of animals sacrifice is distributed equally among dose group. Furthermore, diurnal variations in rats aged 40 days or less seems lower than in adult animals at least for serum TSH in male rats (Jordan et al. 1987; Döhler et al. 1979).

Döhler, K.D., Wong, C.C. & von zur Mühlen, A., 1979. The rat as model for the study of drug effects on thyroid function: consideration of methodological problems. Pharmacology & therapeutics. Part B: General & systematic pharmacology, 5(1-3), pp.305–18.

Jordan, D. et al., 1987. Postnatal development of TRH and TSH rhythms in the rat. Hormone research, 27(4), pp.216–24.

b. An additional source of variation is introduced by the distribution of PND 13 across dose groups as a result of the study design. The breeding phase in these studies is up to 2 weeks in duration, with most occurring during the first 3-4 days. However, mating may not be uniform across dose groups and sampling on any given PND 13 may not include cohorts from all dose groups. Thus control for variation across sampling days may also be a concern.

c. A further confounding factor is that, unlike other studies that involve reproduction and littering, (e.g. OECD TG 416: Two-generation reproductive toxicology study), in the proposed revisions to OECD 421 and 422, litters are not culled to a standard size (e.g. OECD TG 416 recommends culling to 4 males + 4 females). Variations in litter size between dams is normal; for example the litter size for the Sprague Dawley strain of rat varies between 10 – 17 (http://www.criver.com/files/pdfs/rms/cd/rm_rm_r_tox_studies_crlcd_br_rat.aspx).
Variations in litter size are known to significantly affect thyroid hormone levels in both pups and dams (e.g. Strbák *et al*., 1983 and Kahl *et al*., 1991) and so not culling the litters to a standard size will introduce a significant confounding factor when analysis for T_4 and TSH.

d. The review document recommends that dams, but not pups, are fasted for up to 24 h prior to termination and blood collection. This suggestion reflects the fact that the time from last feed to termination can have marked effects on a number of blood parameters, including T_4 and TSH levels. However, fasting is known to have effects on milk output and quality. For example, Viña *et al*., (1986) have demonstrated that even short periods of fasting (6 and 24 h) resulted in marked decreases in amino acid supply and transport in the mammary gland of lactating rats, included a marked decrease in the uptake and transport of tyrosine, which is a key precursor for the synthesis of T_4. Therefore, whilst starvation may help control variation

in dam T_4 and TSH, it represents a significant confounding factor for pup T_4 and TSH. This confounding factor is likely to be further compounded by variations in litter size, with pups from larger litters more likely to be affected as milk output and quality decreases.
We agree that fasting mothers while they give milk is a bad idea and we might consider to not fasting the mothers in these TGs.

Summary and Conclusions

Based on the concerns regarding statistical power and study design, technical and logistical considerations described above, incorporation of T4 and TSH measurements into OECD TGs 421 and 422 as proposed is not appropriate.

References

Kahl *et al*., 1991. Effect of lactational intensity on extrathyroidal 5'-deiodinase activity in rats. *J. Dairy Sci.* **74(3)**: 811-818.

Ottenweller and Hedge, 1982. Diurnal variations of plasma thyrotropin, thyroxine, and triiodothyronine in female rats are phase shifted after inversion of the photoperiod. *Endocrinol.* **111(2)**: 509-514.

Strbák *et al*., 1983. Thyroid hormones in milk: Physiological approach - a review. *Endocrinol. Exp.* 17(3-4)**: 219-235.**

Viña *et al*., 1986. Effects of fasting on amino acid metabolism by lactating mammary gland: Studies in women and rats. *J. Nutr.* **117(3)**: 533-538.