Table 1 Response to comments on MALE PUBERTAL ASSAY

Comment ID Number	Reviewer	Comment	Response
1. Clarity of purpose of the assay			
1.1	RD	The first paragraph of the stated purpose of the assay is clear enough but could be improved by eliminating or replacing some of the phrases. For example, the phrase "information that will be useful in assessing the potential of a chemical substance or mixture to interact with the endocrine system" is much too long and passive. Consider replacing it with "information useful in determining the potential of chemicals or mixtures to interact with the endocrine system.	The suggested change would eliminate the idea that the purpose of this protocol is to obtain information from an *in vivo* system, and in particular a *mammalian* in vivo system. Since these are important parts of the purpose, the statement of purpose is being left unchanged.
1.2	RD	Purpose and Applicability is clearly worded and states the purpose of the assay precisely in such a way that individuals with scientific training can easily comprehend it. It might require re-phrasing for a less technically audience or addition of a lay summary.	Agree. No change in protocol.
1.3	RS	The background information and protocol description give a clear view of the objectives of the assay and the role of its component parts. This material should be easily comprehensible to anyone intending to use and apply this assay.	Agree. No change in protocol.
1.4	TZ	The purpose of the assay is clear. It is difficult to imagine what a novice in this field	Agree. No change in protocol.

		would require to perform the assay as intended by the EPA; presumable, the contract labs performing this would have experience.	
1.5	TZ	The objectives are clearly and concisely articulated in the protocol.	(Protocol is being changed to reflect the purpose statement as written in the Integrated Summary Report since it appears that this is the version on which the other reviewers commented, and is the one which EPA intended to take priority.)
2. Relevance of the assay to its purpose			
2.1	GD	I believe that the biological and toxicological relevance of the assay is well described in section IV (p. 9). I believe that this assay, as well as the adult male and pubertal female assays being evaluated, has the potential to provide the most reliable and comprehensive information for the weight-of-evidence determination. The use of an intact animal model provides the opportunity to assess multiple endocrine processes, both alone and in integration with the hypothalamic-pituitary axes that control thyroid and gonadal function. The ability to measure multiple modes of action in a single assay provides the opportunity to obtain a lot of information from a relatively small number of animals, vs. running separate tests for each mode of action. The intactness of the hypothalamic-pituitary-	Agree. No change in protocol.

		gonadal and hypothalamic-pituitary-thyroidal axes makes the model biologically relevant, as these axes act in concert in the organism that we wish to model for the purposes of hazard and risk assessment, the human. The pubertal male also has special relevance in that puberty represents a major developmental stage in the maturation of the reproductive system, and may be particularly susceptible to exogenous agents that interfere with the hormonal control of sexual maturation. Because young, rapidly growing animals are used, the system is expected to be sensitive to agents that affect thyroid function. Stoker et al. (2000) provides an excellent review of the literature on the effects of exogenous agents on puberty and thyroid function in the pubertal rodent. This review provides a strong theoretical underpinning of the relevance of the assay for the stated purpose of screening for potential effects on androgen and thyroid-dependent systems, and possibly other modes of action.	
2.2	GD	The model is toxicologically relevant because the responses in an intact system, which also has homeostatic mechanisms, is likely to be much more concordant with the results of more definitive toxicity tests.	Agree. No change in protocol.
2.3	RD	In terms of biological relevance, the assay endpoints reflect measures of the integrity of	Agree. No change in protocol.

		the hypothalamic-pituitary- androgen (HPA) and -thyroid (HPT) axes. These include changes in tissue weight, histology, and circulating hormone levels. These endpoints are sensitive to exposure of known androgen and thyroid agonists and antagonists. The endpoints used for the HPT axis are also the most appropriate for the length of the assay. Therefore, the assay measures physiological endpoints appropriate for detecting alterations in the status of the male reproductive and thyroid organ systems. In terms of toxicologic relevance, the endpoints selected for the Pubertal Male Rat Assay are appropriate for several reasons. First, they reflect biologically relevant endpoints as discussed above. Second, validation studies using known androgen receptor agonists and antagonists demonstrate these endpoints are altered by exposure to methyl testosterone, vinclozolin, flutamide, p,p'-DDE and other AR agonist/antagonists. Third, exposure to a dopamine antagonist, pimozide, showed that the assay was sensitive to compounds that inhibit prolactin release. Fourth, exposure of test animals to propylthiouracil, a thyroid hormone synthesis inhibitor, and to phenobarbital, which increases metabolism	

		of thyroid hormones, demonstrated the assay can detect compounds that alter the production or clearance of thyroid hormones. Finally, the endpoints are relevant because competent investigators, whether from industry, contract laboratories or academia are capable of measuring them in a consistent manner.	
2.4	KG	The assay was designed to detect chemicals that interfere with androgen or thyroid function or with the HPG axis based on the understanding of the biological relevance of these functions for normal pubertal development. Serum hormones and reproductive organ weights significantly increase in male rats during puberty and as a result, chemicals that disrupt endocrine function can have a dramatic impact on male pubertal developmental measurements such as organ weights and preputial separation. This assay is highly relevant for toxicological screening for endocrine active chemicals.	Agree. No change in protocol.
2.5	RS	A considerable amount of data has been accrued on this assay and has involved several different (but experienced) laboratories and the testing of a large number of compounds with a wide variety of purported or known mechanisms of action (MOA). The evidence presented for review and in a few publications substantiate the	Agree. No change in protocol.

		view that this assay is fit for purpose.	
2.6	TZ	The assay is relevant to the goals of the EDSP. Data from this assay will not likely provide novel biological information, although it could provide the motivation to address specific mechanistic hypotheses.	Agree. No change in protocol.
3. Transferability of the protocol			
3.1	GD	Transferability study: This study was the first conducted outside of an investigative research lab and confirmed that the protocol yielded similar results in a separate laboratory. The study was extensive and evaluated six chemicals representing different modes of action, in two strains of rats. The results are thoroughly presented, and I agree with the overall interpretation that the protocol is transferable.	Agree. No change in protocol.
3.2	KG	Results from the interlaboratory validation demonstrate that the protocol is transferable and reproducible and capable of detecting chemicals that act through a variety of endocrine related mechanisms to impact male pubertal development.	Agree. No change in protocol.
4. Repeatability and reproducibility of the assay			
a. General comments			
4.a.1	GD	I believe that the results are promising. It is	Agree. No change in protocol.

		possible to run the protocol and obtain reasonably comparable results across laboratories and over time. Different labs were able to detect signals of endocrine activity, and it is unlikely that there would have been many, if any mistakes in the false negative direction had the assay been testing unknowns.	
4.a.2	RD	Based upon the interlaboratory validation and transferability studies, the assay consistently gives results that are both reproducible and repeatable both within given laboratory and between laboratories with minor exceptions due to exceeding the CV for a number of endpoints.	Agree. No change in protocol.
4.a.3	KG	The reproducibility and transferability of the assay is clearly demonstrated by the reproducibility of overall results across laboratories. While there was some variability with some endpoints between the laboratories the overall weight of evidence and conclusions were consistent.	Agree. No change in protocol.
4.a.4	RS	From the validation and other studies plus published studies, the reproducibility of the assay is impressive. From the inter-laboratory study involving two doses of each of four compounds (dibutyl phthalate, vinclozolin, 2-CNB and DE-61), the inter-laboratory reproducibility extended in almost all instances to both doses of each of these compounds. This was all the more	Agree. No change in protocol.

		impressive when considering that in many instances in the same comparison, most or all of the laboratories were unable to meet the CV performance criteria for these endpoints (discussed earlier). This imparts considerable confidence that the assay is inherently robust and reproducible and will be transferable between laboratories with relative ease; the comparatively simple format of the assay components reinforces this conclusion. The only outstanding issue is that of the false positive rate in the assay, but this should be resolvable in time with its wider application to chemicals with unknown activity.	
4.a.5	TZ	"Due to an oversight, serum hormone levels (T4, TSH, testosterone) were not obtained in this study." This demonstrates that "GLP" means only that record keeping is precise, not that the study was performed according to plan, or that the techniques used to perform the study were appropriate or adequate.	The "oversight" was an EPA error in instructions to the laboratory for this particular contract, not in the laboratory's adherence to Good Laboratory Practices.
b. Variability in endpoint values			
4.b.1	GD	The multi-dose study is thoroughly described. There is only one point on which I believe the interpretation should be expanded: p. 36, lines 4-7, it is explained that an apparent result of flutamide on	As shown in the performance criteria section of the Integrated Summary Report (pp. 50-52), the coefficient of variation for organ weights in historical controls has been in the range of 10-20%. For adrenals, it was around 15%. EPA believes that this

		adrenal weight was probably due to inordinately low values in two control animals and was not an effect of the compound. I agree with this conclusion, but it raises the question of whether this apparent variability problem is important enough to correct. For example, in another part of the report, it was determined that high variability in the weights of fluid-filled organs indicated that more detail and training was needed across labs in proper dissection. In this instance, could it also be a problem with procedure, or is the variability due to the animals themselves, in which case it may be important to either increase Ns or to amass a larger historical control data set against which new results can be compared.	is an acceptable variability, whether the variability is due to the technical abilities of the laboratory or to variability in the animals. It does not appear necessary at this time to increase the number of animals used in the assay. The Agency agrees that it would be useful to re-examine historical variability in controls as more data become available.
4.b.2	GD	The extent of variability for many of the endpoints is troubling: all labs were out of compliance with pre-set performance criteria for 4 of 17 endpoints for one lab, 5 of 17 for two, and 6 of 17 for one. In other words, roughly one-fourth to one-third of the endpoints were more variable than was believed to be acceptable, a result that could compromise the resolving ability of the assay (as well as its reproducibility). These are issues that will need to be addressed in order for the assay to be used routinely to evaluate unknowns.	The Agency agrees that the performance of the laboratories in the interlaboratory validation study was outside of the historical norms for a surprising number of endpoints. However, it would not be appropriate to assume that these laboratories reflect the variability likely to be encountered during the Screening Program better than the laboratories from which the performance criteria were derived. Due to the redundancy of many of the endpoints in the assay, the laboratories were able to provide consistent assessments of the ability of the compounds to interact with the endocrine system (the goal of Tier 1

			of the Screening Program) despite the variability in some of the individual endpoints. For this reason, the Agency believes it would be inappropriate to withhold this assay from use in the battery in order to determine whether compliance with the performance criteria is a serious problem. The Agency considered loosening the performance standards in order to accommodate more of the values encountered in the validation study but did not do so because it would reduce the confidence in the endpoint-specific information that might help indicate mode of action.
4.b.3	RD	The only consistent endpoint that was exceeded [sic] the CV was determination of ventral prostate weight. These exceptions were mainly due to dissection technique differences. However, failure to keep the CV within the stated range did not prevent determination of an effect.	Agree. No change in protocol.
4.b.4	KG	There were inconsistencies in hormonal measurements between laboratories. This is likely due to biological variability but may also have to do with technique. Despite the inconsistencies the overall trend was consistent across laboratories and the redundancy of endpoints reduces concern regarding any one specific measurement. Thus, while there is some variability associated with specific endpoints in this assay, the inclusion of multiple endpoints	Agree. No change in protocol.

		increases its reliability.	
4.b.5	RS	Arguably one of the more variable and more subjective aspects of the pubertal assay is the endpoint of preputial separation (PPS). This is clearly a useful 'endocrine' endpoint that summates androgen action over a period of time, as shown in all of the studies so far done to test and hone the pubertal assay. Nevertheless, in the inter-laboratory study none of the 4 (experienced) labs involved could meet the coefficient of variation (CV) criterion for bodyweight at PPS and only 2 could do so for age at PPS. Some refinement of the PPS 'definition' was adopted after issues relating to incomplete PPS (retention of 'threads' of connecting tissue), but by the very nature of the assessment it seems to me that PPS will always be prone to high between-laboratory and between-observer variation. Similar to PPS, the recorded weights of the ventral prostate, seminal vesicles + coagulating gland, epididymis and levator ani +bulbocavernosus muscle all provide a measure of androgen action over time – in essence they summate androgen action over the 30-day course of the assay. Accordingly, weights of these organs changed more or less in parallel to PPS in response to exposure to the various compounds tested in the different parts of	Agree. No change in protocol.

		the validation exercise and in other studies. My initial reaction to this (expected) observation was that PPS was maybe redundant, and could therefore be dispensed with, as it did not measure anything that the target organs already did not. However, it is also apparent that for most of these organ weights there is similarly high CV as discerned from the inter-laboratory study. Thus, none out of the 4 labs met the CV criterion for ventral prostate weight, only 1 met the CV for seminal vesicle weight and only 2 did so for epididymal weight. None of this is unduly surprising, as anyone experienced in the dissection of these organs will know that not only is there high variability in actual weight (probably largely reflecting different levels of fluid content), but the dissection process can also be variable depending on how this is done. For this reason it is good practice to have the same person do all of the dissections for the same organ in order that variation within a laboratory is minimized. Returning to the issue of whether PPS is worth retaining, I convinced myself that it was, based on two lines of reasoning. First, it is an 'in-life' measurement, and thus may provide the first indication of 'anti-androgenic' or 'androgenic' activity in a test	

		compound which can then be confirmed by organ weight measurements. Second, as PPS and reproductive organ weights are all intrinsically highly variable measurements (for reasons outlined above), it is safer to have multiple endpoints that reflect the same underlying phenomenon/activity (ie. androgen action over time), as this will increase the chances of detecting a significant effect on any one of the endpoints; the fact that one is in-life and the others terminal reinforces this argument. Additional to this reasoning is that PPS is non-invasive and not time-consuming as the visual inspection can be made at the same time as dosing the animal.	
4.b.6	RS	Essentially two analytical methods are used as part of the test, hormone assays and selective evaluation of organ histopathology (testes, epididymides, thyroid, kidney). For the most part, the hormone measurements do not constitute an important component of the pubertal rat assay, but if these are to be retained as part of the overall assay then standardization of the assay kits used is essential to provide uniformity as well as minimizing inter-laboratory variation. However, such variation is commonplace and likely to be considerable when, and if, the pubertal rat assay is put into widespread use by laboratories that have little	The Agency believes that the hormone measurements add value to the assay and should be retained despite the interlaboratory variation that has been observed. There are instances in which a chemical can cause changes in T4 without concomitant changes in thyroid histopathology, for example, making this hormone measurement useful in detecting interaction with the thyroid system. The performance criteria and the discussion of requirements for use of hormonal assays that is included in the protocol may help keep variability manageable. However, the Agency agrees that the hormone assays are not the strongest endpoints in the assay, and emphasizes that the amphibian metamorphosis assay is important for the evaluation of the ability of a chemical to interact with

		experience with running hormone assays.	the thyroid system.
4.b.7	TZ	The RIA data provided in this document show a great deal of variability in hormone levels of the control animals across laboratories. However, it is not possible to identify the source of this variation as being technical or biological because the types of studies required to separate these two sources of variation were not performed. Specifically, the EPA should develop and distribute, or should contract to develop and distribute, the quality control standards to all laboratories performing RIAs in the commission of the EDSP. These centralized standards would greatly decrease the variance across laboratories and would enhance the reliability of the assays. In addition, the three laboratories used different commercial kits for the various RIAs and EPA did not require that the RIAs were validated (in the case of heterologous assays) or that the QC was performed as described by the kit manufacturer or that the performance fell within the range defined by the manufacturer. There is no question that these problems can account for a great deal of variability in the RIA results, and that a minimal amount of thought and effort by the EPA at the beginning of this project could have prevented it. It must be remembered	The protocol (section X, Hormonal Assays) will be changed to read as follows: "Hormonal measurements can be conducted using radioimmunoassay (RIA), immunoradiometric assay (IRMA), enzyme-linked immunosorbent assay (ELISA), or time-resolved immunofluorescent procedures. Regardless of which is used, always include multiple quality control (QC) samples run in duplicates that are dispersed among the test samples. Any measurement kit that is used must be shown to yield appropriate values for control rats at the laboratory performing the pubertal assay. This includes demonstrating that QC was performed as described by the kit manufacturer and that the performance falls within the range defined by the manufacturer. The lab's criteria for evaluating the kit's performance must be included in the study report. If the laboratory has never had experience with the kit for making measurements specifically in the rat, it should test the kit in one or more untreated rats outside of the pubertal assay before relying on it for the full study." The Agency cannot commit at this time to being the source for quality control standards for all laboratories performing RIAs in the commission of the EDSP. Use of historical quality control samples maintained in-house, and/or use of such samples from the manufacturer, is judged to be sufficient for the

		that RIAs have been in use for nearly 50 years, and methods for validating assays and standardizing them across laboratories have been very well developed.	purposes of the EDSP. See also the response to comment 7.4.
4.b.8	TZ	An important question is whether the variability in endpoints can be reasonably reduced – both within lab and between labs – by standardizing different elements of the test. A major variable will be that of the feed. We know that variation in the amount of estrogenic compounds in feed is high, regardless of the supplier's certification. This variation alone can interact with test compounds to provide variability from lab to lab, or at different times within the same lab. Variability in hormone levels will be affected by this, but also by the standardization methods as described above.	The Agency understands this issue of standardizing all of the potential variables in the assay but feels that the additional research necessary to completely understand the effects of each of the variables and their interactions would considerably delay implementation of the Screening Program. The assay was demonstrated to be sensitive to known endocrine-active agents, and the Agency believes it prudent to begin screening chemicals using the current form of this assay rather than wait for further optimization and revalidation. As for the specific issue of estrogenic compounds in feed, the Agency believes that there is no evidence that variation below the cap placed on total genistein equivalents will affect the endpoints in this assay, and that there is some evidence, though sparse, that it will not affect them. In addition, the requirement in the protocol that the same batch(es) of feed must be used for both controls and treated animals will minimize the variability between labs or at different times with the same lab. Finally, the concern for estrogens in feed is of less concern in the male pubertal assay than in the female pubertal assay.
c. Dose selection			

4.c.1	GD	There were differences among labs in the dose levels at which some effects were detected, which may have an influence on assay performance if dose selection is not perfect.	The purpose of this assay is to detect interaction of a chemical with the endocrine system, not to determine the lowest dose at which effects might occur. While there might be some differences among labs in sensitivity, these are minimized by requiring testing at the maximum tolerated dose.
5. Clarity of protocol			
5.1	GD	Clarity and conciseness of the protocol in describing the methodology of the assay such that the laboratory can: a. comprehend the objective This section is succinct and to the point. b. conduct the assay The protocol contains enough information for a competent laboratory to conduct the assay in a consistent way. There are a few aspects of the protocol that may be too restrictive, such as the admonition to keep temperature and relative humidity within ranges that may not be achievable in all facilities, and are different from those established by AAALAC. c. observe and measure prescribed endpoints The information in the protocol and attachments were clear and helpful in providing guidance on evaluating the endpoints. I found it very useful that the attachments to the protocol included	The protocol is being changed to allow a wider range of temperature and humidity (viz., the range specified in DHHS/PHS NIH Publication No. 86-23, 1985, *Guidelines for the care and use of laboratory animals)*.

		information useful to conducting the assay, such as reference images for thyroid histology. d. compile and prepare data for statistical analysis Sufficient guidance was provided on how the data should be displayed and analyzed. The statistical procedures were not overly restrictive but provided enough guidance to facilitate comparison of study results in different labs. e. report results The protocol provided a great deal of information on how to report and interpret results. The length of the data interpretation section is unusual in my experience, but given the novelty of the protocol and the complexity of interpretation, I believe it to be warranted.	
5.2	RD	The instructions on how to conduct the assay are complete and clear for the most part.	Agree. No change in protocol.
5.3	RD	observe and measure prescribed endpoints- clear and concise	Agree. No change in protocol.
5.4	RD	compile and prepare data for statistical analyses- clear and concise but consider specifying statistical software.	The Agency does not believe that specifying statistical software will materially improve the analysis of the data inasmuch as the analyses required by the assay are not extraordinary or unusually complex.
5.5	KG	The protocol is clear and comprehensive. The objective is clearly stated and sufficient	Agree. No change in protocol.

		detail is presented to allow a laboratory with the appropriate expertise to conduct the assay and accurately analyze and report the results. Methods for housing and treatment of the animals are presented in sufficient detail. Each endpoint is clearly described and methods for statistical analysis as well as how to handle outliers are presented. Finally details and examples for data interpretations, presentation, and developing a final report are given.	
5.6	RS	Insofar as I feel competent to judge (as a scientist running an academic research laboratory), the protocol provided is clearly laid out, is understandable and is sufficiently detailed to enable an appropriately experienced laboratory to run, complete, evaluate and report results using this assay. There are no deficits in the protocol that I have noticed.	Agree. No change in protocol.
5.7	TZ	In general, this is a well-written protocol and a well-written and well-managed validation study.	Agree. No change in protocol.
5.8	TZ	To the best of my ability to determine, the protocol is sufficiently detailed that an experienced laboratory could conduct the assay as written.	Agree. No change in protocol.
5.9	TZ	Generally, the prescribed endpoints are clearly articulated.	Agree. No change in protocol.
5.10	TZ	The protocol is clear in directing laboratories in their data preparation and analysis.	Agree. No change in protocol.

6. Clarity of reporting format			
	RD	report results- clear and concise.	Agree. No change in protocol.
6.1	TZ	Likewise, the kind of information requested in the report from studies are clearly represented in the protocol.	Agree. No change in protocol.
7. Performance criteria			
7.1	GD	The final paragraph on p. 72 indicates that, while the assay results are generally reproducible, the CV criteria that were established a priori were not always met. I don't see this as a fatal flaw with the assay, but it will be important in the early going to constantly re-evaluate the magnitude and sources of variability and to find ways to minimize the latter and better set acceptable criteria for the former.	The Agency agrees that the magnitude and sources of variability need to be examined carefully as data from the Screening Program come in, and that it may be necessary to adjust the performance criteria in the future.
7.2	TZ	Clearly, the EPA thought about performance criteria and about the logic required to interpret the findings. The performance criteria should be sharpened, both or the way the endpoints at necropsy are evaluated as the RIA performance criteria [sic].	See responses to comments 7.4 (concerning RIA performance criteria) and 9.b.4 (concerning histopathology).
7.3	TZ	The performance standards for the RIAs does not take into consideration that the contract labs are reporting performance of	The reviewer appears to be referring to RIA values obtained from the interlaboratory validation study. These values were not used in setting the

		their assay that falls outside the performance standards reported by the manufacturer.	performance standards.
7.4	TZ	In addition, the EPA has developed performance standards, which will likely improve the quality of the data received from this assay. However, the mechanism by which these performance standards are generated should be more closely evaluated. Moreover, there are no performance standards established for the RIAs. These commercial RIA kits come with manufacturer-established performance characteristics, but the EPA does not require that contract labs use these kits in a manner that is consistent with the manufacturer performance. Finally, some of the kits being used in this assay are heterologous (i.e., prepared and calibrated for human samples, but used in rats), and the EPA does not require the contract lab to validate the assay. Given this situation, it is no wonder that there is a high degree of variation in hormone measurements. It is highly likely that the variability in hormone levels observed in these experiments can be reduced to such an extent that hormone levels themselves can play a larger role in the in vivo portion of the EDSP.	As described in the response to comment 4.b.7, the following wording is being added to Section X, Hormonal Assays: "This includes demonstrating that QC was performed as described by the kit manufacturer and that the performance falls within the range defined by the manufacturer." The Agency believes that the T_4 assay developed for humans is relevant for use with rats. As explained in more depth in the response to peer review comments on the adult male assay, the Agency examined the percent recovery across multiple concentrations of T_4 in the pool of rat serum reported in the results by RTI and Charles River. The recoveries ranged from 90 to 110%. Additional evidence that supports the use of the thyroid hormonal assay kits with rat serum was the T_4 results following phenobarbital treatment in the adult male assay. Absolute and relative changes were highly consistent with the results in the vehicle-control group and treatment groups across laboratories for each dose level and in accord with toxicological and biological historical results (Adult Male ISR Summary Tables 9 and 14, respectively). Furthermore, the manufacturer of the T_4 assay kits provided technical in-house data where they spiked rat serum with multiple concentrations of T_4 and reported similar ranges in percent recovery. Hence, the results in the ISR combined with the results

			reported in two review articles and veterinary application documents by Diagnostic Products Corporation (DPC) referenced below support the general consensus that commercial T_4 assay kits developed for use with human serum are relevant for use with rat serum. Davies DT. 1993. Assessment of rodent thyroid endocrinology: Advantages and pit-falls. Comparative Haematology International 3:142-152. Christian MS, Trenton NA. 2003. Evaluation of thyroid function in neonatal and adult rats: The neglected endocrine mode of action. Pure and Applied Chemistry 75:2055-2068. Siemens Medical Solutions Diagnostics (Diagnostic Products Corporation, DPC), Coat-A-Count TKT3 (total T_3) and TKT4 (total T_4), Veterinary application documents of T_3 (March 26, 1993, ZV106 A) and T_4 (March 24, 1993, ZV103 A). The high variability of these labs can be from many sources (which are out of EPA control), including animal handling, necropsy methods to control stress, the duration of the necropsy (increased can increase variability due to the diurnal pattern of most hormones), and lastly the RIA itself (which can have factors within which will increase variability-from something as simple as pipetting the iodinated hormone at the same time, within minutes, or if they

			incubated according to manufacturer's instructions).
8. Data interpretation			
8.1	GD	First, on p. 52, lines 14-18, it is stated that EPA does not require the assay to consistently display a pattern of endpoint responses diagnostic for a particular mode of action, but only that thyroid responses not be used to claim consistency with sex steroid associated responses and vice versa. Given that the stated purpose of the assay is to detect a variety of hormone-related modes of action, and that, as the most apical assay in the screening battery, it will have the greatest influence on weight of evidence determination of the battery, we need to expect more of the assay.	The Agency agrees that it would be desirable to have better consistency of diagnostic patterns across laboratories than was obtained in the interlaboratory validation study. However, the assay was shown to be able to detect interaction with the endocrine system even if the mode of action cannot always be determined. Since the goal of Tier 1 screening is to detect interaction, not mode, the Agency believes it appropriate to include this assay in the Screening Program now rather than to wait until greater consistency of diagnostic patterns can be ensured.
8.2	RD	For the multiple chemical studies, the effects of phenobarbital exposure on the thyroid axis were noted as non-significant but the trends were in the correct direction. One reason suggested for the lack of significance was failure to reach the MTD. Failure to reach the MTD may be common when there is a lack of data concerning the toxicity of test compounds. Perhaps a quantitative method of determining significance when a number of endpoints almost reach the stated level of significance could be used or the standard could be multiple endpoints	It is unlikely that the phenobarbital study as run would be accepted by the Agency as an adequate study if it had been submitted as part of the Screening Program, in part because the MTD was not reached, and in part because the thyroid hormone measurements were not made. The Agency believes that reaching the MTD and doing a complete study are critical in order to conclude that a test substance does not interact with the endocrine system. It would be inappropriate to lower the standards for significance as a substitute for reaching the MTD. Developing a method for consistently evaluating the

		reaching the 0.10 level of significance as a way of determining weight of evidence.	weight of evidence when MTD has been reached, none of the endpoints reach statistical significance, but trends are seen may be appropriate and Agency will consider this suggestion.
8.3	RS	The interpretation of the results obtained using this assay in the different laboratories, including several inter-laboratory comparisons, are rational and fit with current understanding of how the various endocrine systems operate within the body during pubertal development. There are some issues in relation to assay specificity and which endpoints are absolutely essential, and others which might be dispensable or assigned a 'supporting role' (see below), but these are rather minor issues, and they do not affect the overall conclusion that the assay appears robust and fit for purpose, but with some limitations.	Agree. No change in protocol.
8.4	RS	In the validation studies, organ histopathology proved to be a rather insensitive endpoint as effects were only detected for compounds fully expected to have major target organ effects, namely DBP and 2-CNB on the testis and propylthiouracil, perchlorate and DE-71 on the thyroid. As each of these compounds had corresponding effects on organ weights and/or on relevant hormone levels, a case could be made for dispensing with organ histopathology in this Tier 1 assay,	As the Agency evaluates the data that come in from the Screening Program, it will consider whether the histopathology information adds enough useful information to warrant continued inclusion. The kidney histopathology is included as a marker of general toxicity, and is expected to be useful in determining whether the MTD has been exceeded.

		especially as it requires considerable histopathological expertise. Even though I am a great fan of histopathology, I am not convinced that it adds greatly to the assay, bearing in mind its primary objectives. I see no need for kidney histopathology.	
8.5	RS	As far as I am able to judge, the statistical methods used for analysis of the significance of effects, for analysis of trends and for comparison of variability in methodology between laboratories is appropriate. However, I am not sure that any statistical package can truly evaluate the performance of the assay as this has to integrate all of the organ and hormonal data in a way that allows objective decision making and classification and I am not certain that this is possible. Instead, I feel that such decision making will be based not on appropriate individual statistical tests but analysis of the data by experts who have experience of the test and with results and variability in responses that it shows for different chemicals.	Agree. No change in protocol.
8.6	TZ	"…The conclusion from this study was that the pubertal male assay clearly identified atrazine as interacting with the endocrine system at both dose levels, thus showing that the assay is sensitive to chemicals that affect the HPG axis…". EPA's conclusion is dangerous! Rather, this study shows that	The reviewer may have misunderstood how the assay will be used in the battery and what the intent of the validation effort was. Atrazine was tested as a known positive, just like the other chemicals aimed at disrupting receptor binding or steroidogenesis. When used in the battery, given what we know about this chemical, we would have made the same conclusion

		when the EPA has previous information indicating that a chemical interacts with the endocrine system, they can selectively interpret the data in a way that is consistent with what is known. It is important to recognize that this entire paragraph amounts to arm waving. It would be interesting to see what would happen if the EPA were to test chemicals in a blinded study in which neither the contract lab, nor the EPA interpreters were aware of the test chemical.	(i.e., it is a centrally acting compound) because it does not bind to ER or AR or affect steroidogenesis and that an effect was observed on pubertal development in both sexes. We are showing that atrazine would be detected in the male and female pubertal assays proposed by EDSP (which would be relevant for chemicals with similar targets), as it was shown to delay VO and PPS. Though one may not be able to pinpoint the exact mechanism of the chemical within the brain/pituitary/gonadal axis, when reviewed in light of the other information that would be obtained in the battery, this conclusion is sound.
8.7	TZ	Page 26, lines 9-14 [of the Integrated Summary Report]. This conclusion ignores the observation made directly above it that serum T4 levels were reduced, although serum TSH was not altered, nor were there treatment-related histological changes in the thyroid gland. No basis is given for ignoring T4 levels.	T_4 changes would need to be interpreted with the other four thyroid endpoints (TSH, thyroid weight, histology and liver weight) in addition to body weight and other EDSP assays, before we would trust a single endpoint.
8.8	TZ	Page 28, lines 14-18 [of the Integrated Summary Report]. This paragraph illustrates a weakness in the EPA's logic. First, low thyroid hormone can cause a decrease in weight gain; thus, animals treated with high levels of PTU could have a lower body weight precisely because serum T4 levels are lower. In addition, serum T4 levels are sensitive to caloric restriction; therefore, if animals treated with a high dose of compound such that caloric intake is	The restricted feeding study, cited in the Integrated Summary Report, showed that reduced body weight gain up to 6% did not interfere with the thyroid endpoints of the pubertal assay, and the Data Interpretation Procedure therefore cautions that "…studies which suggest thyroid activity only at a dose level causing more than approximately 6% body weight loss at termination compared to controls may need to be repeated at a lower dose level." The Agency also emphasizes that the male pubertal

		restricted, serum T4 levels could be lower due to this mechanism.	assay is not intended to stand alone in the determination of interaction with the thyroid system. The amphibian metamorphosis assay will provide information that may be more convincing and reliable than the information from the male pubertal assay for thyroid effects.
8.9	TZ	"Thyroid weight was increased at both doses, though the increases were not statistically significant (27.3 mg in controls, 31.9 and 32.5 mg at the low and high dose levels, respectively). As a fundamental rule, if something is not statistically significantly different, it is not different. Also as a fundamental rule, the biological significance of a difference may be arguable, but if an endpoint is statistically significantly different, it should be reported and interpreted as such; and if not….	Agree. No change in protocol.
8.10	TZ	"The conclusion of this part of the study was that the male pubertal protocol appears to be sensitive to the thyroid-related and gonadal effects of Phenobarbital even though the thyroid-related responses were not significant at the $p<0.05$ level." This exact same profile was observed with Linuron and with Flutamide, yet the EPA concluded that the thyroid endpoints were not important. These studies are showing only that EPA can identify a well-known endocrine disruptor, not that they can identify an endocrine disruptor for which	The Agency agrees that phenobarbital was not positive in this study, but notes also that the dose level tested was not as high as it should have been in a correct application of this assay. The wording "appears to be sensitive … even though the…responses were not significant at the $p<0.05$ level" was intended to mean that the results were not inconsistent with a positive result that might have been seen had the dose level been at the appropriate level, even though the results at this dose level were negative. The negative finding in this study for phenobarbital, of

		there are no previous data.	course, does not show that the assay *is* sensitive for thyroid effects when run correctly. This supports the need to include another assay for thyroid effects in the Tier 1 Battery – for example, the amphibian metamorphosis assay.
8.11	GD	The conclusion about phenobarbital (p. 32, lines 27-29) is that the protocol was sensitive to the thyroid-related effects of Phenobarbital despite the fact that there were no statistically significant changes in most of the thyroid-related responses. The report blames the lack of significance on the fact that the MTD wasn't reached and that therefore the experiment was not an adequate test of the resolving power of the assay. I can accept that rationale, but if true, then nothing can be said about the ability of the assay to detect Phenobarbital as an indirect thyroid toxicant. The preceding text on p. 32 is clear about that point, but it should be reflected in the concluding paragraph.	See response to comment 8.10.
8.12	TZ	A significant weakness is that the interpretation of the thyroid endpoints seems to require previous knowledge of the activity of the test compound. The justification of this statement is that the profile of effects observed with Phenobarbital was identical to that observed with other compounds like Linuron and Flutamide, yet the interpretation was based on previous publications.	See response to comments 8.10 and 4.b.10.

		Moreover, all hormone levels are not being measured in a manner that will lead to logical interpretation.	
9. Appropriateness and completeness of validation			
9.a. Number and kinds of chemicals tested in validation			
9.a.1	GD	There is not enough information to assess the specificity of the assay. It may be too non-specific to provide enough information to correctly classify negatives. More research needs to be done to address this concern	The Agency made every attempt to identify a chemical that was considered to be systemically toxic but also demonstrated to be without effects on the endocrine system in both the male and female. There were, of course, many chemicals which are known to be systemically toxic but which have not been reliably tested for endocrine-related effects, so these could not be used as indicators of specificity. Chemicals have been identified that were endocrine active in the male but not the female or vice versa, or which affect the thyroid system but not the reproductive axis. These chemicals support the idea that there is specificity. The Agency will examine the data on the first set of chemicals to go through the Screening Program carefully for evidence of specificity or lack thereof.
9.a.2	GD	There is not enough information yet to make definitive conclusions about the range of modes of action that are detectable by the assay. The evidence is good that the assay can detect anti-androgens and thyroid-active agents. There is also reasonable evidence that the assay can detect androgens, agents that affect the hypothalamic-pituitary gonadal axis, and steroidogenesis inhibitors.	The Agency agrees that the evidence for the sensitivity of this assay to estrogens and aromatase inhibitors is mixed, and that this argues for inclusion of assays in the Tier 1 battery that provide additional information on these modes of action. The Agency removed the claim that the male pubertal assay is responsive to estrogens when it prepared the Integrated Summary Report.

		The evidence on estrogens and aromatase inhibitors is mixed. These may not be limitations of the assay, but are limitations on how we can interpret the utility of the assay at present. Clarity on which modes of action it covers will be important in determining the assays that should be conducted with it as part of a battery.	
9.a.3	GD	Fourteen chemicals were evaluated during the course of assay development and validation. The chemicals represent a range of modes of action, including androgens, anti-androgens, estrogens, thyroid-active compounds, and agents that affect the hypothalamic-pituitary-gonadal axis. In some instances chemicals with the same mode of action but different potencies were evaluated. The test substances were appropriate to evaluate the range of mode of actions which the assay is capable of detecting. Given the number of potential modes of action it would be very useful to have a much larger set of chemicals evaluated. However, I am aware of the number of other tasks that EPA needed to perform to validate other assays in the battery, and feel that the chemicals in the validation set for the pubertal male assay are representative of a wide range of modes of action and provided a rigorous test of assay performance. Only one presumably	Agree. No change in protocol.

		negative substance, chloronitrobenzene, was evaluated, and this compound had effects on assay endpoints. It will be important to evaluate assay performance with additional negative compounds.	
9.a.4	RD	- a wide ranging selection of test compounds was selected that included antiandrogens that ranged from weak to strong, compounds that had varied mechanisms of action on the androgen-pituitary-hypothalamus axis and the thyroid synthesis and metabolic pathways.	Agree. No change in protocol.
9.a.5	KG	A weakness of this assay may be its ability to screen for week estrogens. While the assay can detect more potent estrogens, there is insufficient data regarding weaker estrogens.	Before taking this assay to peer review, the Agency deleted the claim that this assay could detect estrogens.
9.a.6	KG	In addition, the lack of a test of a negative control makes it difficult to determine the specificity of this assay for endocrine active compounds.	See the response to comment 9.a.1.
9.a.7	KG	2-CNB was chosen as a toxic compound that did not affect endocrine function. Unfortunately, this compound demonstrated endocrine disrupting activity and did not serve as a good negative control.	Agree. No change in protocol.
9.a.8	RS	Although results obtained in applying the assay to a variety of chemicals with known endocrine activity have been largely as expected, the specificity of the assay is to some extent unproven. In the inter-	Agree. No change in protocol. See also the response to comment 9.a.1. The Agency is considering additional research on 2-chloronitrobenzene to determine a potential mode of

		laboratory study, 2-chloro nitrobenzene (2-CNB) was included as a test of specificity as various indicators had suggested that it would not have 'endocrine activity' and would thus provide a test of assay specificity. However, 2-CNB significantly reduced testosterone levels, weights of all androgen-dependent organs and delayed preputial separation as well as causing histopathological changes to the testes. Based on these findings, 2-CNB would clearly be classed as an endocrine disruptor and it must be presumed that this is the case (nitrobenzene and dinitrobenzene are well-established testicular toxicants in the adult rat, although I am not aware of evidence that they disrupt Leydig cell function). The alternative interpretation is that the pubertal assay is prone to non-specific effects and will therefore yield 'false positives' with a high frequency. Although it is preferable for a Tier-1 (screening) assay to suffer from this rather than from frequent false negatives, it will be important in future studies to more rigorously investigate the specificity of the pubertal assay. My expectation is that the assay will not be unduly prone to false positives, based on the relative lack of impact of the feed restriction studies on the assay endpoints, but this needs to be demonstrated	action for the results observed in this study. This may help distinguish between endocrine-specific effects and other toxicity which might lead to "false positives".

		categorically using appropriate test compounds.	
9.a.9	RS	The validation process, which involved testing of a considerable number of compounds with a range of known activities and mechanisms of action provided solid foundations for subsequent evaluation and interpretation of results for compounds of unknown hormone activity. It also identified unexpected effects or results, such as those for 2-CNB and phenobarbital, which are discussed elsewhere. It is likely that continued application of the assay to a wider range of chemicals will uncover other activity profiles that do not fit within our expected concepts, but this will inevitably lead to a better understanding of the utility, and limitations, of the assay.	Agree. No change in protocol.
9.a.10	RS	The validation studies have used compounds with a wide range of hormonal or other activities and these have provided a robust evaluation of the effectiveness of the assay and of its sensitivity and discriminatory powers. This data has been obtained from several different but experienced laboratories. The main criticism is that no truly negative compound was evaluated; 2-CNB was chosen for this purpose, although I am frankly amazed that anyone would choose such a compound as a 'negative' when closely related chemical	See the response to comment 9.a.1. As for the choice of 2-CNB despite its similarity to profound testicular toxicants, note that the Agency made an effort to find chemicals for which reproductive and/or developmental effects were looked for but not found, in the belief that such chemicals were the most reasonable candidates for being endocrinologically inactive. Most chemicals in the NIH database for reproductive and developmental toxicity apparently were tested because of their similarity to chemicals known to have such effects. This was the only chemical (besides polymers, etc.)

		compounds have been shown to be profound testicular toxicants in adult rats. Aside from this, the range of compounds chosen was wise and included a notably wide range of different hormonal activities. This choice undoubtedly proved that the assay is readily able to detect compounds that affect the 'androgenic' hormonal systems at various levels, and data for the thyroid axis (in which I am inexpert) also appear reasonably convincing, with the notable exception of phenobarbital effects.	that appeared to have no reproductive or developmental toxicity, but did show minimal other toxicity indicating that a sufficiently high dose was tested. It may also be relevant that Blackburn et al (Tox Appl Pharm 92:54-64. 1988) showed testicular toxicity only from the 1,3 isomer of dinitrobenzene. The 1,2- and 1,4- isomers were without effect on the testis.
9.a.11	TZ	The test substances were appropriately chosen as chemicals that produced a relatively weak effect. This was a good test of the sensitivity of the assay and resulted in somewhat ambiguous results with Phenobarbital.	Agree. No change in protocol.
9.b Analytical methods used in validation			
9.b.1	GD	The analytical methods were appropriate. Most of the endpoints are of organ weight and histology, which were straightforward, if somewhat variable for some tissues (e.g., ventral prostate). I believe that the variability will decrease as labs become more adept in the dissections and standardized in their procedures. The hormone measurements were conducted according to accepted methodology.	Agree. No change in protocol.

9.b.2	RD	analytical methods- highly appropriate foe endpoints utilized in the assay.	Agree. No change in protocol.
9.b.3	KG	The analytical and statistical methods appear appropriate.	Agree. No change in protocol.
9.b.4	TZ	The analytical methods were variable. Certainly body and organ weights are reasonably analyzed. However, histopathological analysis should be objectified with computer-assisted morphometry. Morphometry has been shown to be more reliable than visual estimation (1). In addition, the RIAs are problematic as has been discussed above.	Computer-assisted morphometry was judged to be too expensive for this screening assay. See the response to comment 7.4 concerning RIAs.
9.c Statistical methods used in validation			
	GD	The analysis of interlab performance included a calculation of each lab's CV for the endpoints measured, which was an appropriate way of identifying robust endpoints and potential areas for improvement in assay performance. The graphical summaries of the CVs across laboratories in the interlab reproducibility study were useful in understanding the range of variability in the study. These analyses appear to have been appropriately done.	Agree. No change in protocol.
	RD	statistical methods in terms of demonstrating the performance of the	Agree. No change in protocol.

		assay- Not my area of expertise. I consult with a statistician to select appropriate methodology but the methods listed are the ones recommended to me for similar studies.	
	KG	The analytical and statistical methods appear appropriate.	Agree. No change in protocol.
	TZ	The description of the statistical methods is clear and appropriate.	Agree. No change in protocol.
10. Special issue: thyroid			
10.1	GD	The second result that deserves some additional discussion is the number of compounds that affect thyroid hormone (T4) concentrations. It appears that T4 is not a particularly specific measure of thyroid toxicity. This lack of specificity has been noted in other screening protocols (e.g., O'Connor et al., 2002) and should be discussed in this report in some depth, either in this section or in the overall conclusions. It is possible that T4 measurements can only be interpreted in the context of some other effect to be meaningful. The more that this is discussed before this or other assays come on line, the better the interpretation of results will be for compounds with uncharacterized modes of action.	The Agency agrees that data interpretation for effects on the thyroid system is strongest when there are other endpoints (including endpoints in other relevant assays) than just T_4. The male pubertal assay includes thyroid weight, thyroid histology, and TSH level as additional thyroid-related endpoints. The amphibian metamorphosis assay and the female pubertal assay are expected to provide information useful in a weight-of-evidence assessment of effects on the thyroid system. Unfortunately, due to the complexity of the thyroid system there does not appear to be an in vitro assay (or a reasonably limited set of in vitro assays) that would provide appropriate information that could replace the use of these in vivo assays.

10.2	RS	I was somewhat disconcerted to see that, in both the multi-chemical study and in the TherImmune2 study, phenobarbital was detected as a clear anti-androgen, based on delayed PPS and reduced reproductive organ weights; equally disconcerting was its lack of significant effect on the thyroid axis. My presumption was that phenobarbital had been included as a test compound primarily because of its potential thyroidal effects (clearly evident in the adult male rat assay) and because its central effects might also 'spill over' into effects on the reproductive axis. As there is published data to show that phenobarbital can suppress LH secretion and even inhibit (fetal testis) steroidogenesis, the effects in the pubertal assay are not entirely unexpected but I was surprised at their robustness. It is possible that the pubertal rat is especially vulnerable to central effects of phenobarbital as normal progression of puberty is dependent on a progressive increase in frequency and amplitude of LH pulses, which then drive increasing testosterone production; the alternative interpretation, namely that the pubertal rat is more susceptible to non-specific effects remains to be resolved, as discussed above.	The Agency agrees that the non-thyroid effects of phenobarbital were robust in this assay and regards this as additional support for the usefulness of this assay to detect HPG-axis disruptors. The lack of any observed change in thyroid hormone measures was likely due to the dose of phenobarbital employed. Again, these data indicate the need to properly employ an MTD in dose selection and emphasizes the need to interpret the male pubertal assay in light of the data from the other thyroid-sensitive assays to be included in the Tier 1 Battery.
10.3	TZ	"At this point, no environmental chemicals have been found to bind to the thyroid	The male pubertal assay is expected to detect agents that bind to the thyroid hormone receptor: TSH would

		receptor. (See Stoker et al. (2000b) for review of toxicant effects on thyroid function.)" This statement is incorrect. Several compounds have now been shown to bind to the TR, some with IC50's in the nM range. The authors are not even using the EPA-managed thyroid DRP.	decline and T_4 would increase. Changes in thyroid weight and histology are also expected.
11. Strengths of the assay			
11.1	GD	a. The assay is performed in an intact system, capable of responding to interactions among multiple endocrine axes.	Agree. No change in protocol.
11.2	GD	The assay uses developing animals, which may be more susceptible to endocrine perturbations, thereby potentially increasing the sensitivity of the assay.	Agree. No change in protocol.
11.3	GD	c. The assay measures endpoints indicative of a number of modes of action of interest to EPA for endocrine screening. It should be possible to obtain a lot of information out of a relatively few animals.	Agree. No change in protocol.
11.4	GD	d. The apical nature of the assay will allow it to provide a lot of context for weight-of-evidence interpretation of some of the simpler assays in the proposed tier 1 battery. It also allows for multiple endpoints for each mode of action, improving the interpretability of the assay.	Agree. No change in protocol.
11.5	RD	Strengths of the assay include ease of	Agree. No change in protocol.

		conducting the assay and measuring the endpoints, the moderate duration of exposure, biological and toxicological relevance and substantial literature base. In addition, this assay uses intact animals whereas other assays such as the Hershberger assay require castrated animals. Furthermore, many of the techniques used are common to other assays such as the 15-day adult rat assay, multigenerational testing protocols, and Hershberger assays that have been used for some time. The use of time of PPS as an endpoint is an advantage due to its increased sensitivity to androgen agonists and antagonists compared to weight of androgen-responsive organs and tissues. Coupling this assay with the pubertal female rat assay provides a robust test capable of detecting xenoestrogens, antiestrogens, androgen agonists and antagonists as well as inhibitors of thyroid synthesis, inducers of thyroid hormone metabolism and compounds that alter release of pituitary hormones.	
11.6	KG	Strengths of this assay include the ability to screen for multiple modes of action in a sensitive in vivo mammalian assay. The assay has extensive historical data from multiple laboratories and the biology behind the various endpoints is well understood.	Agree. No change in protocol.

		The assay focuses on a time period when reproductive organ development is very sensitive to endocrine disruption. Because it is in vivo the assay allows for consideration of absorption, distribution, metabolism, and excretion. The assay is relatively rapid and has been standardized so that it can be performed in any laboratory that has the appropriate expertise and experience. The assay has multiple and redundant sensitive endpoints that can be used to help design more definitive Tier-2 testing. For example, results with 2-CNB provided sufficient information to suggest that its effect on the growth of androgen dependent tissues is either through altering steroidogenesis or targeting the secretion of pituitary hormones but not through interference with the androgen receptor.	
11.7	RS	It is a Tier-1 assay and its priority is to maximize the detection of endocrine active compounds (that target the male reproductive and/or thyroid axes) whilst minimizing false negatives. The use of multiple endpoints (mainly organ weights) is designed to ensure this.	Agree. No change in protocol.
11.8	RS	The use of an intact animal with normally functioning, homeostatic, hormone systems is, in my opinion, a strength as it represents the ‘real world’. It includes metabolism etc and can potentially ‘integrate’ chemical	Agree. No change in protocol.

		effects that may directly affect one hormonal (or non-hormonal) system and then secondarily affect another hormonal system.	
11.9	RS	As with the intact adult rat assay, the pubertal rat assay uses multiple endpoints in order to maximize detection of compounds with weak activity or with a profile of activity that does not fit within expected boundaries (for example a compound that exhibits both anti-androgenic and anti-thyroidal activity). It is simpler than the adult male assay in terms of endpoints (less hormonal data), and arguably this makes the assay potentially more straightforward to operate and interpret, though it requires a considerably longer treatment period. The use of multiple endpoints may provide preliminary information on the potential MOA but, in my opinion, the main importance of the inclusion of multiple endpoints in this Tier-1 assay is to maximize the likelihood of detection of endocrine active chemicals whilst minimizing the chance of false negatives.	Agree. No change in protocol.
11.10	RS	I anticipate a progressive ability to categorize chemicals into classes based on their activity profile in this assay, even if it is not possible to define a clear MOA.	Agree. No change in protocol.
11.11	RS	Because the test uses an intact animal that is advancing through one of the most endocrinologically dynamic phase of its life	Agree. No change in protocol.

		(puberty), it might be anticipated that compounds might affect target organs or hormone levels via pathways that are unrelated to endocrine disruption *per se*, for example effects on food intake/metabolism that lead secondarily to such changes. However, from results obtained so far, and including food restriction studies, this expectation has not surfaced in any major way, which arguably makes the assay more robust, simpler in its function and analysis and relatively easy to interpret.	
11.12	RS	The fact that only 3 hormonal assays are included, namely for thyroxine (T_4), thyroid stimulating hormone (TSH) and testosterone, makes the pubertal assay relatively easy to run, as hormone assays are often a source of considerable inter-laboratory variation and, because they provide only a 'snap-shot' in time of hormone levels, they can be potentially misleading.	Agree. No change in protocol.
11.3	RS	The four main strengths of the assay are: (1) that it has a strong foundation based on various studies that have been undertaken as part of the validation exercise using a variety of compounds with different activities; (2) the primary reliance on target organ weights, which are easily measured, as the most predictive endpoints; (3) the use of multiple endpoints to inform on the same	Agree. No change in protocol.

		'hormonal activity', some of which extend beyond organ weights (histopathology, hormone levels) and one of which is a dynamic pubertal 'in-life' event measured over time (PPS); (4) the minimal use of hormone measurements, as these are not only more technically difficult to measure, but may be difficult to interpret at a time in development when there can be large fluctuations within, and especially between, animals.	
11.14	TZ	The strength of the assay is that it evaluates the interaction of toxicants with the androgen and thyroid systems during a particularly sensitive period of development for these interactions. Thus, this assay should be more sensitive that the adult in identifying EDCs. The endpoints for androgen action are valid and known to be sensitive to changes in hormone action during this period.	Agree. No change in protocol.
12. Limitations of the assay			
12.1	GD	The assay is lengthy. Dosing covers a 30 day period. … The length of the assay is not conducive for rapid screening.	The original protocol was shorter (20 days of dosing) but did not go much past the normal age of puberty to assess delays in preputial separation. The additional 10 days increase the possibility of detecting effects of endocrine-active chemicals on pubertal development. In addition, we have shown that the extended duration of exposure enhances the ability to detect thyroid

			alterations compared to the 20 days in the female pubertal protocol. Gray et al., 2002, Xenoendocrine disrupters-tiered screening and testing, filling key data gaps. Toxicology 181-182, p. 371 -382.
12.2	GD	Furthermore, the study requires a pilot study for dose-setting.	Dose-setting for this assay is no different in concept than dose-setting for other assays and is not considered a part of the assay itself. While it is true that there is often more data relevant to dose-setting for assays in adult systems than for an assay using this period of development, when such data are not available for an adult animal, it too must be developed.
12.3	GD	The assay appears to be inordinately influenced by changes in body weight gain, such that a significant change in body weight gain is needed for a valid (presumably, sensitive) assay, but changes above 9-10% make the assay difficult to interpret. It may not be possible to achieve this much precision in dose setting, especially in a screening context with limited numbers of animals.	There is usually little variability in body weight gain within a group, particularly when a group is standardized at weaning by body weight as is required by the protocol. This should ease the problem of setting the highest dose level, although the Agency agrees that choosing levels for the dose-setting study can be difficult.
12.4	GD	The apparent non-specificity of T4 is concerning, and may suggest that this measure cannot be interpreted out of context with other thyroid measures.	See response to comment 10.1.
12.5	RD	Limitations of the assay are its inability to determine more downstream effects such as sperm production, motility and fecundity.	Agree that these endpoints are more appropriate for Tier 2. No change in protocol.

		However, assays that detect these endpoints are more appropriate for Tier-2.	
12.6	RD	The intra- and interlaboratory variability in the hormone assays make it more difficult to detect subtle changes with any degree of significance.	See response to comment 10.1.
12.7	RD	This assay does not differentiate between the various mechanisms of action by which a compound can affect androgen status or whether change in thyroxine is due to inhibition of synthesis or induction of metabolism.	The purpose of the assay is to identify the potential for interaction with the endocrine system, not to determine mode of action. It is an added benefit if, sometimes, it is possible to get accurate information that is useful in determining mode.
12.8	RD	This assay also uses timed pregnant rats which are not only live animals but fairly expensive.	Determination of the day of birth is extremely important for successful use of this assay (particularly body weight and puberty endpoints), and is known to be of inadequate reliability when supplied by commercial vendors. Thus it is necessary to observe the day of birth as part of the assay. Also, it is not clear that the cost of the relatively few pregnant dams that are needed for this study is greater than the cost of the relatively numerous offspring that would need to be purchased.
12.9	RD	Although the exposure period is short in comparison to some other assays, it does require 30 days.	See response to comment 12.1.
12.10	RD	Last, this assay is more recent than the Hershberger assay or the rat uterotrophic assays which translates into a smaller data base and less harmonization.	While the database for the pubertal assay may be smaller than for the uterotrophic or Hershberger assay, the assay still appears to be relevant and repeatable. It is relevant to a wider range of endocrine modes than those assays, and thus is not strictly comparable.

12.11	KG	The reliance on MTD is a weakness since additional prior studies must be performed to accurately identify MTD or in cases where the MTD is based on a review of the literature may lead to over or underestimating the MTD.	See response to comments 12.2 and 12.3.
12.12	RS	The main weakness of the assay, which at this stage of evaluation is more theoretical than experience-based, is that exposure of rats to compounds during such a hugely dynamic phase of growth and reproductive development as puberty is likely to result occasionally in effects on reproductive 'endocrine systems' that are not due to intrinsic 'endocrine activity' but to an indirect effect eg. on growth. Although the feed restriction studies did not produce evidence to suggest major impact of this on reproductive organ weights or PPS in the pubertal assay, I still consider this to be a likely event with some compounds; the effects of phenobarbital are worth discussing in this respect (see below). Nevertheless, such an outcome will mean at the worst that some compounds are identified as 'false positives' which is acceptable in any screening assay provided this is not unduly frequent; the issue of assay specificity, which remains to be resolved, is important in this regard and has been discussed above.	The Agency agrees that specificity has not been shown yet, but notes that specificity appears not to be testable at this time since there are no chemicals which have been reliably tested for endocrine effects and shown to be negative. The presence of other toxicities *per se* does not necessarily mean that endocrine effects are not present at lower levels. The Agency agrees that it will be important to evaluate specificity as more data become available.

12.13	TZ	There are no endpoints of thyroid hormone action that are equivalent to those of the androgen system (e.g., secondary sex organ weight). This produces an unbalanced assay that has confounded the interpretation of results of this assay and the Adult Intact Male 15-day Assay. Specifically, several compounds cause a reduction in serum total T_4, including Linuron, PBDEs, PCBs, Phenobarbital. In each case, the interpretation is that these compounds activate liver enzymes (e.g., UDPGTs) that decrease circulating half-life for serum T_4. However, in the case of Phenobarbital (mostly), serum TSH is increased in response to low T_4. In contrast, serum TSH is not always elevated in response to low T_4 produced by Linuron, or PCBs. It is not clear why the same level of serum T_4 is not always associated with an increase in T_4, but this apparently is the case. The question is whether low serum T_4 produces adverse effects on peripheral tissues – especially during development – in the absence of increased TSH. Failure to identify and incorporate valid endpoints of TH action that would be equivalent to (e.g.,) seminal vesicle weight for androgens, represents a significant weakness in the EDSP that will create considerable debate about the interpretation of data derived from	The amphibian metamorphosis assay includes functional endpoints (developmental stage and morphology). The male pubertal assay should not be the only thyroid-related assay in the Tier 1 Battery.

		these assays	
13. Details of the protocol			
a. Strains and species			
13.a.1	GD	The transferability study is also noteworthy in that it adequately addresses the question of strain sensitivity and supports the decision to use Sprague Dawley rats in subsequent studies.	Agree. No change in protocol.
13.a.2	RD	There is no strain of rat listed in this brief description of the assay; shouldn't it be stated here?	The preferred strains are listed in the protocol.
b. Wording			
13.b.1	GD	The section describing the study on the effects of body weight gain is satisfactorily described. There is one semantic correction that I would like to see made in this section, which is to change all references to "body weight loss" to "decreased body weight gain". The latter is actually what was measured. The animals on restricted feed were not losing weight; they were gaining weight at a slower pace than controls. This distinction is important because a misunderstanding of this point could lead to inappropriate dose selection, excessive toxicity, and an uninterpretable study.	While this comment applied to the Integrated Summary Report, not the protocol, the wording in the protocol that refers to "reduction in body weight [of treated animals] compared to controls" will be changed to refer to "reduction in body weight gain [of treated animals] compared to controls."

13.b.2	RD	Replace the words kill, killed or killing with euthanized or euthanasia.	"Euthanasia" is reserved for situations where an animal is killed to terminate its misery. Thus, "kill" is more appropriate when referring to normal termination at the end of the study.
c. Dose selection			
13.c.1	GD	The final paragraph of this section (p. 47-48) needs to be clarified or expanded on. The first sentence of the paragraph states that a 10% reduction in body weight is a reasonable basis for setting an upper limit in the assay. However, the next sentence suggests that this level of effect is too severe for thyroid effects and the final sentence indicates that a 9-10% decrease may necessitate conducting additional studies and/or a weight-of-evidence approach to interpret the results for the thyroid endpoints. This seems to indicate that the male pubertal assay cannot be optimized for thyroid endpoints and reproductive endpoints, and that either a compromise would need to be made in selecting an upper dose level, or an additional assay would have to be run. This is an important point in determining whether the assay is a suitable alternative, and should be discussed at greater length.	As noted in the response to comment 12.3, the Agency agrees that dose-setting may be difficult even though the tightness of the variability in body weights (if that is what MTD is being based on) will help. The availability of the amphibian metamorphosis assay in the Tier 1 Battery is also important to ensure that thyroid effects are covered appropriately.
d. Solvent			

13.d.1	RD	In the second paragraph, it is stated that the test compounds will be dissolved in corn oil whereas test substances are dissolved in a methyl cellulose vehicle for the 15-day adult male assay. Why the difference in choice of vehicles? Corn oil contains varying amounts of phytoestrogens that may mask some changes in the endocrine system. It seems like a more inert vehicle such as triolean would be more appropriate. … Section VI discusses the properties of corn oil to be used but does so in qualitative terms. Consider changing it to corn oil from an approved source, from a freshly opened container, and free of sediment. However, the use of corn oil is problematic in that it can contain phytoestrogens. Why not use an synthetic oil such as triolean?	One vehicle is not likely to be appropriate for all chemicals, so flexibility in choice of vehicle is essential. The wording will be changed to the following: "The test substance is dissolved or suspended in a suitable vehicle. Consideration should be given to the following characteristics: Effects on the absorption, distribution, metabolism, or retention of the test substance; effects on the chemical properties of the test substance which may alter its toxic characteristics; and effects on the food or water consumption or the nutritional status of the animals. Use of vehicles with potential intrinsic toxicity should be avoided (e.g., acetone, DMSO). If corn oil is used, it must be clear and free of sediment. It should have a bland odor, free from rancid, musty, metallic, putrid or any other undesirable odor. Other solvents such as water or carboxymethylcellulose may be used where appropriate. If the test substance is not soluble in any of the conventional solvents, it is administered as a suspension. It is important that the dosing solution or suspension be well-mixed to keep the chemical well-distributed prior to and throughout dosing, and care must be taken to ensure that the particle size of insoluble substances does not interfere with delivery of the full dose through the gavage tube or needle tip."
e. Husbandry other than diet			
13.e.1	RD	In addition, pups should be culled to 8 or 10 per litter rather than the range of 8-10 and	Protocol will be changed to reflect the intent that all cages must have the same number of animals. That

		housed 2 or 3 per cage instead of 2-3 to further reduce variability.	number may be either 2 or 3. (In the case of 2 per cage and a planned N of 15, it will be necessary to add an extra rat to the last cage.)
13.e.2	RD	Section III states that the litters will be standardized or culled to 8-10 pups per litter. For a particular study, all litters should be standardized to the same size.	Protocol will be changed to reflect the intent that all litters in a particular study will be standardized to the same size. That size may be 8, 9, or 10 pups.
13.e.3	RD	Prohibition of the use of tap water for the animals is very good to eliminate perchlorates and other contaminants.	Agree. No change in protocol.
13.e.4	RD	The humidity conditions are perhaps too stringent and may prevent the assay from being performed in regions of the country that are too dry or too humid. For example, facilities in West Texas, New Mexico and Arizona frequently can not maintain 30% humidity.	The protocol will be changed to the following: "Animal care and housing should be in accordance with the recommendations contained in the *Guide for the Care and Use of Laboratory Animals* (Institute for Laboratory Animal Resources, National Research Council. National Academy Press, 1996). In that Guide, the following statement appears: "the acceptable range of relative humidity is 30 to 70%."
13.e.5	RD	Why is it not acceptable to cross-foster to raise the litter size to the minimum? It can prevent waste of animals if used properly.	Inclusion of both the male and the female pubertal assays in the Tier 1 Battery should allow efficient use of as many of the pups in a litter as possible. In addition, cross-fostering would introduce another variable, whose effect on the assay is unknown. Dams have been known to reject offspring that are not their own.
13.e.6	RD	Heat-treated aspen shavings are preferred to heat-treated pine shavings since they contain lower levels of volatile compounds and appear to be less allergenic to workers.	The protocol will be changed to read as follows: "Rats are housed in clear plastic cages (approximately 20 x 25 x 47 cm) with heat-treated laboratory-grade wood shavings other than cedar as bedding. Corn cob bedding should not be used due to its potential to

			disrupt endocrine activity . Wire-mesh-bottomed caging should not be used due to the potential for pup loss."
f. Diet			
13.f.1	RD	Specifying rat chow that is low in phytoestrogens is good but consider adding as an appendix a list of acceptable rodent chows.	Phytoestrogen content is variable even within a specific type of chow, so listing "acceptable" chows would not be possible. Measurement of the phytoestrogen content of each batch used in a study is important. Some chows are marketed as "soy- and alfalfa-free", and these may be appropriate sources to investigate, although other chows may meet the criterion as well.
g. Method of kill/euthanasia			
13.g.1	RD	Section IX needs to be revised to meet AVMA Panel on Euthanasia standards.	The 2007 AVMA Guidelines on Euthanasia (http://www.avma.org/issues/animal_welfare/euthanasia.pdf), which replace the 2000 Guidelines, state "[Decapitation] is conditionally acceptable if performed correctly, and it should be used in research settings when its use is required by the experimental design and approved by the Institutional Animal Care and Use Committee." Since testosterone level is an endpoint in this assay and is affected by acute stress, the experimental design requires the use of a rapid method such as decapitation.
13.g.2	RD	Decapitation without prior use of either CO2 or inhalational anesthetic is not an AVMA Panel on Euthanasia (2000) approved	See response to comment 13.g.1. The Agency could find no requirement in the AVMA Guidelines for peer-reviewed journal articles that conclude prior exposure

		method of euthanasia. It is listed as a conditionally acceptable method by the Panel but requires justification for use such as peer-reviewed journal articles that conclude prior exposure to CO2 or inhalational anesthetic alters the endpoints that are the focus of the study. The statement is made in the protocol that decapitation is considered more humane than CO2 asphyxiation. This is in direct opposition to the AVMA Panel on Euthanasia which is reference used by IACUCs, AAALAC, the NIH Guide, and the USDA.	to CO_2 or inhalational anesthetic alters the endpoints that are the focus of the study in order to justify use of decapitation. Similarly, the Agency could find no comparison of the humaneness of decapitation vs. CO_2 asphyxiation in the AVMA Guidelines. The Agency's statement that it finds decapitation more humane than CO_2 asphyxiation is based on Holson R. 1992, Euthanasia by decapitation: evidence that this technique produces prompt, painless unconsciousness in laboratory rodents. Neurotoxicol Teratol 14(4):253-257 when compared to the statement in the AVMA Guidelines that "High concentrations of CO_2 may be distressful to some animals."