SUPPORT PROTOCOLS

No. 3. Propagation

1. Remove the medium from the culture dish with sterile pipette or sucker.
2. Rinse the cell with 5 mL of PBS.
3. Remove PBS with sterile pipette or sucker.
4. Add 2 mL of Trypsin-EDTA solution (0.25% Trypsin + 0.02%EDTA/PBS) to cover the bottom of the culture dish and then remove the excess.
5. Allow to stand Trypsin treated cell for ca. 3 min in 5% CO_2 incubator at 37°C.
   (Monitor cells under microscope. Cells are beginning to detach when they appear rounded)
6. Tap the dish gently.
7. Wash to remove the adherent cells with 5 mL of 10%FBS-EMEM.
8. Count cell number.
9. Dilute the cell suspension with 10%FBS-EMEM to 0.4-1.0 x 10^5 cells/mL.
10. Place 10 mL of cell suspension to 90 mm culture dish.
11. Incubate the cell in 5% CO_2 incubator at 37°C.

6/8/2005

SUPPORT PROTOCOLS

No. 4. Preparation of frozen stock

1. Remove the medium from the culture dish with sterile pipette or sucker.
2. Rinse the cell with 5 mL of PBS.
3. Remove PBS with sterile pipette or sucker.
4. Add 2 mL of Trypsin-EDTA solution to cover the bottom of the culture dish and then remove the excess.
5. Allow to stand Trypsin treated cell for ca. 3 min in 5% CO_2 incubator at 37°C.
   (Monitor cells under microscope. Cells are beginning to detach when they appear rounded)
6. Tap the dish gently.
7. Wash to remove the adherent cells with 5 mL of 10%FBS-EMEM.
8. Count cell number.
9. Centrifuge the tube at 1100 rpm (200-300 x g) for 5min, and remove the supernatant carefully.
10. Add Cell-Banker* (Juji Field Inc.) and resuspend the cell at density of ca 1 x 10^4 cells/mL.
11. Make 1 mL aliquots of cell stock.
12. Freeze and store the cell stock below -80°C**.

*Conventional freeze medium (90% FBS/10% DMSO) can be used in place of Cell-Banker.
**Storage in liquid nitrogen would be preferable for long-term storage (more than 3 months).

SUPPORT PROTOCOLS

No. 5 Preparation of assay plate

Prepare a dish of cultured hERα-HeLa-9903 cell

1. Remove the medium from the culture dish with sterile pipette or sucker.
2. Rinse the cell with 5 mL of PBS.
3. Remove PBS with sterile pipette or sucker.
4. Add 2 mL of Trypsin-EDTA solution to cover the bottom of the culture dish and then remove the excess.
5. Allow to stand Trypsin treated cell for ca. 3 min in 5% CO_2 incubator at 37°C.
   (Monitor cells under microscope. Cells are beginning to detach when they appear rounded)
6. Tap the dish gently.
7. Wash to remove the adherent cells with 5 mL of 10%FBS-EMEM and transfer the cell suspension to a centrifuge tube.
8. Count cell number.
9. Centrifuge the tube at 1100 rpm (200-300 x g) for 5min, and remove the supernatant carefully.
10. Resuspend the cell with 10%FBS-EMEM to obtain a final cell density of 1 x 10^5 cells/mL.
11. Add 100 μL of cell suspension into each well of 96 well assay plate (Nunc #136102 or equivalents).
12. Incubate the cell in 5% CO_2 incubator at 37°C for 3h
13. Proceed to chemical exposure.

SUPPORT PROTOCOLS

No. 6-1. Chemiluminescence Detection with standard luciferase reagent

Reagents

Cell lysis reagent (4.5x): Dilute 10 mL of 5×Cell Culture Lysis Reagent (CCLR, #E1531) with 45 mL of distilled water.

Luciferase Assay Reagent: Add 1 vial (105 mL) of Luciferase Assay buffer (Promega, #E4550) into a vial containing Luciferase Assay Substrate (Promega, #E4550), and dissolve the substrate thoroughly. Store the substrate below -20°C if necessary.

Chemiluminescence Detection

1. Flick and drain off the contents of the assay plate.
2. Add 100 μl of PBS to the well to wash the plate.
3. Flick and drain off the contents of the assay plate.
4. Add 100μl of PBS to the well to wash the plate again.
5. Flick and drain off the contents of the assay plate.
6. Add 15 μL of Cell lysis reagent (4.5x) to wells.
7. Incubate for 10 min at room temperature.
8. Add 50μL of Luciferase Assay Reagent to wells.
9. Read plates on a Chemiluminescence plate reader.

SUPPORT PROTOCOLS

No. 6-2. Chemiluminescence Detection with luciferase reagent using Steady-Glo Luciferase Assay System

Reagents

Luciferase Assay Reagent: Add 1 vial (100 mL) of Luciferase Assay buffer into a vial containing Luciferase Assay Substrate (Promega, #E2520), and dissolve the substrate thoroughly. Store the substrate below -20°C if necessary.

Chemiluminescence Detection

1. Remove 50 μL of assay medium from all wells of assay plate.
2. Add 100 μL of Luciferase Assay Reagent to wells.
3. Allowed to stand for 5 min.
4. Read plates on a Chemiluminescence plate reader

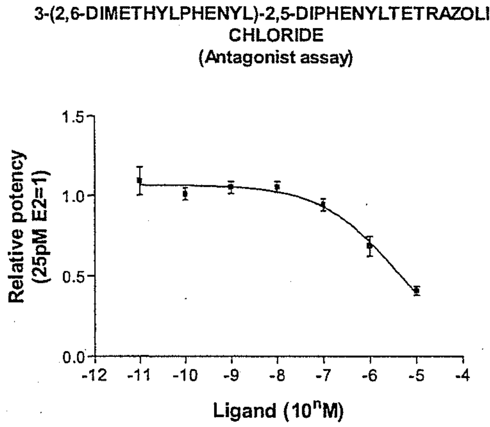
Monitoring of cytotoxic effect of chemicals in reporter gene assay

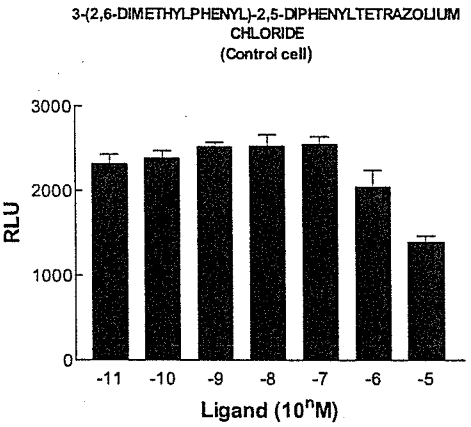
March 15, 2006
Masahiro Takeyoshi, CERI-Japan

Cytotoxicity is the quality of being toxic to cells caused by toxic agents (chemical substance). In general, cytotoxicity can be measured by the MTT assay or other conventional methods (Alamer dye method etc.). Reporter gene assay is an analysis method that allows the identification of promoters and enhancers and the study of the correlations between their activities and conformations by checking the amount of the reporter proteins that are expressed from reporter genes. And the endpoint of hER-HeLa-9903 cell based reporter gene assay is a luciferase activity that is produced as a result of the transcriptional activation of the reporter gene. Cytotoxic effect of chemicals may lead misunderstanding of the results of this assay system, especially in reporter gene assay for antagonist activity of chemicals.

In our system, cytotoxicity detection system using control cell, which constantly produces firefly luciferase by the RSV promoter without any stimulation, is already established for antagonist assay system (Please refer to the document entitled "***Outline of ERα Antagonist assay using hER-HeLa-9903***" dated March 15, 2006).

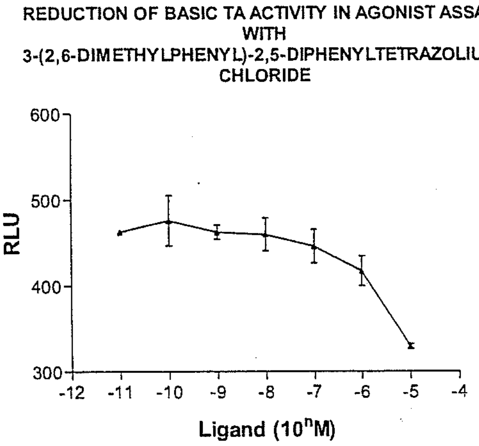
In this system, cytotoxicity of chemical is clearly detectable as shown below;

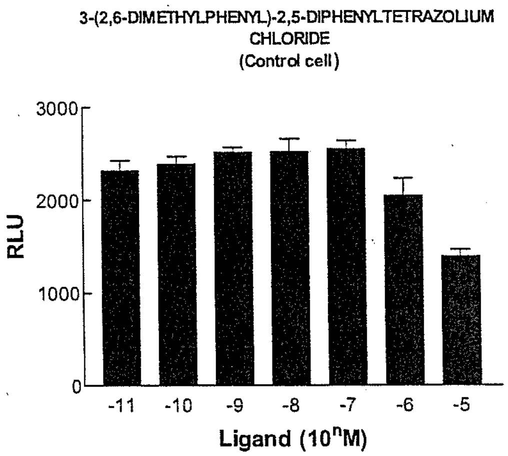




In this case, the antagonist like effect observed in the antagonist assay is concluded as negative because of its cytotoxicity.

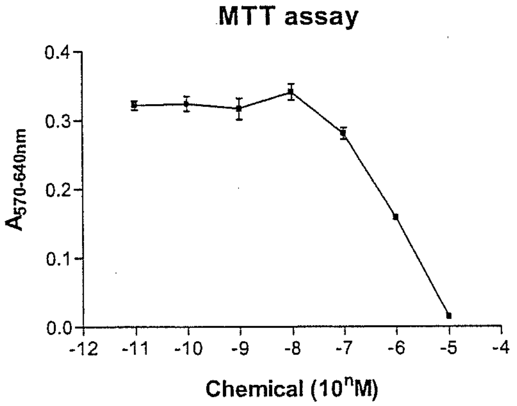
The cytotoxic effect of chemical also causes reduction of basic transcriptional activity in agonist assay (See below).

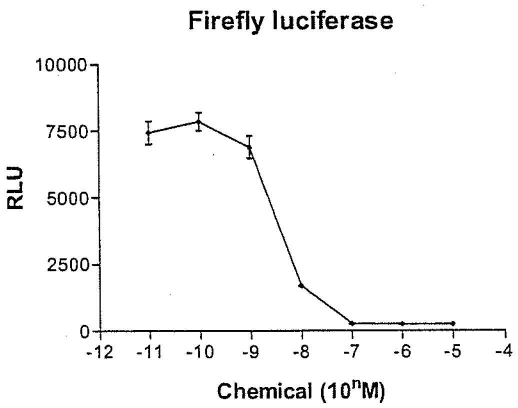

REDUCTION OF BASIC TA ACTIVITY IN AGONIST ASSAY WITH 3-(2,6-DIMETHYLPHENYL)-2,5-DIPHENYLTETRAZOLIUM CHLORIDE


3-(2,6-DIMETHYLPHENYL)-2,5-DIPHENYLTETRAZOLIUM CHLORIDE (Control cell)

This result indicates that monitoring of basic TA activity in agonist assay can provide the cytotoxic effects of chemical.

In some laboratory, MTT assay may be employed for monitoring cytotoxic effect of chemicals. MTT assay is a general experimental technique for measuring cellular proliferation (cell growth). In this assay, the amount of yellow MTT (3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) oxidised to purple formazan is measured spectrophotometrically. This oxidation takes place when mitochondrial reductase enzymes are active, and thus conversion is directly related to the number of viable cells, another way of saying it is related to the number of cells possessing active mitochondrial reductase enzymes.
However, an endpoint of the reporter gene assay is luciferase activity resulting from the transcriptional activation, and is not a mitochondrial reductase activity.
For this reason, clear discrepancy is noted between the cytotoxicity measured by MTT and that monitored by luciferase activity.
Figures below shows cytotoxic effects of tripropyl-tin measured by both two methods.


MTT assay


Firefly luciferase

Although the transcriptional activity measured by luciferase, this means cytotoxic effect on cellular transcriptional activity, was definitely reduced at least 10^{-8}M of tri-propyl tin, no effect was noted in MTT assay at the same concentration (10^{-8}M of tri-propyl tin).
This suggests that cytotoxicity in the reporter gene assay should be monitored with luciferase activity of control cell or basic transcriptional activity of agonist assay rather than MTT assay.

Appendix 2

Appendix 3

Subject: FW: Statistical approach for intra- and interlaboratory variability

------ Forwarded Message
From: <Hasemanjk@aol.com>
Date: Thu, 25 May 2006 10:56:50 -0400
To: "Deal, Frank H (NIH/NIEHS) [C]" <dealf@niehs.nih.gov>
Cc: "Tice, Raymond (NIH/NIEHS) [E]" <tice@niehs.nih.gov>, "Ceger, Patricia (NIH/NIEHS) [C]" <cegerp@niehs.nih.gov>, "Blackard, Brad (NIH/NIEHS) [C]" <blackard@niehs.nih.gov>, "Charles, Jeffrey (NIH/NIEHS) [C]" <CharlesJ2@niehs.nih.gov>
Conversation: Statistical approach for intra- and interlaboratory variability
Subject: Re: Statistical approach for intra- and interlaboratory variability

Frank-

I have examined Dr. Aoki's PowerPoint slides, and I believe I understand his concerns.

The examples used to illustrate his concerns involve data from four labs with three runs per lab. These 12 data points (logPC50s) are apparently each based on estimates from a Hill equation analysis. However, regardless of how the estimates are obtained, each of the logPC50s is an estimate and has an associated SE of the estimate. One of Dr. Aoki's objections is that these standard errors associated with the estimation process are typically ignored in the data evaluation process.

For example, the typical approach for computing the mean response for each lab is to simply average the three runs. Dr. Aoki prefers instead a weighted average approach that weights each estimate inversely with the associated variability (i.e., the less variable estimate gets weighted more heavily in the averaging process). In my opinion, this is a reasonable option, and I suspect that a statistical purist would likely prefer the weighted average approach to the unweighted average. However, it could also be argued that since each run was carried out under identical conditions, the runs should be given equal weight, regardless of variability.

Thus, I disagree with Dr. Aoki that it is 'naïve' and 'inappropriate' to work with unweighted means, which provide unbiased estimates of the underlying parameter and typically are similar to the weighted means in any case. For example, in one of Dr. Aoki's examples, the unweighted mean is -6.94; the weighted mean is -6.93. I suspect that this is typical of what would be found in practice, especially since there are 'validity check' safeguards built in that will minimize the likelihood that the underlying variability estimates will differ greatly from run to run. From a practical point of view, it is unlikely in our area of application that the choice of weighted vs. unweighted means will have any noticeable impact on the overall interpretation of a study.

I note also that in Dr. Aoki's Slide 6, the lab and run columns are mislabelled and should be reversed.

A second related concern of Dr. Aoki is the calculation of an SD. For example, the variation in response among the three runs at a given lab in theory represents two distinct sources of variability: (i) the variability associated with the estimation process itself; and (ii) the additional variability that might be due to factors that are different from run to run. The SD that is normally calculated does not distinguish between these two sources of variability, but Dr. Aoki feels that this distinction is important and that by subtracting out (i) and focusing strictly on (ii), one obtains better 'estimates'.

Better estimates of what? I agree that his approach provides better estimates of Source of Variability (ii), but I would argue that the primary variability of interest is the actual observed variability among runs, which reflects both (i) and (ii). It should not matter if this variability is due entirely to the estimation process (as was the case in three of the four labs in his example) or if both (i) and (ii) contribute to this variability. The end result is what matters.

Similar comments apply when combining the lab means to produce an overall average. Once again, one could either use a weighted average (-7.15 in Dr. Aoki's example) or an unweighted average (-7.13). Generally, the two will agree very closely.

The variability observed among the lab means is due to a combination of three sources of variability: (i) and (ii) as noted above and (iii) additional variability introduced by factors that differ among labs. Here again, Dr. Aoki recommends 'subtracting out' (i) and (ii) to obtain a 'pure' estimate of (iii). I would once again argue that it is the overall variability that is important, regardless of the contribution of the three individual components.

Although weighted versus unweighted means will very likely have little or no impact on the final interpretation of a study, the same may not be true for an evaluation of variability. In Dr. Aoki's 'fake data' example, he concludes that the much better SD's are essentially all zero. What does this mean from the standpoint of assessing the reproducibility of the assay? I worry that a naive investigator may assume that this means that the assay is extremely reproducible (after all, it has zero SD's), but this may not be the case at all. It may simply mean that the variability associated with the estimation process is so great that it can totally account for the overall variability in response observed among runs and among labs. The magnitude of this variability may or may not be cause for concern, but I still would argue that quantifying the specific sources of the variability is not nearly as important as evaluating the magnitude of the resulting variability itself, as assessed in the 'traditional' (and not 'inappropriate') way.

Dr. Aoki states in Slide 15 that the statistical programs used to produce the Hill equation estimates of the logEC50 do not provide associated SE estimates, but I do not believe that this is the case. Doesn't Prism produce them routinely? If so, then this information can be used in the manner suggested by Dr. Aoki.

Importantly, in the final analysis, one must decide if the purpose of these studies is to refine our estimates of the various sources of variability that contribute to differences in response, or is it to determine whether or not an assay has acceptable reproducibility. Dr. Aoki's presentation focuses on the former, but in my opinion, the latter should be our goal. Thus, if I am trying to determine whether or not an assay is acceptably reproducible, I would want to focus on the observed variability in the actual EC50 estimates across and within labs regardless of the factors that contributed to the variability.

For example, suppose I observed a coefficient of variation of 50%, that in normal circumstances would be unacceptable. However, using Dr. Aoki's approach, it is not this variability that is important, but the relative contribution of the factors that produced it.

This high variability might be due to the estimation process, differences among runs, differences among labs, or a combination of these three factors. In my opinion, quantifying these sources of variability and determining which is the primary contributor should not be our focus. For example, one extreme possibility is that the Hill equation model fit is so poor (and the resulting SE's of the estimated EC50's so high) that Source of Variability (i) can account for essentially all the variability in response, and as a result all the better estimate SD's computed by Dr. Aoki for Sources of Variability (ii) and (iii) are close to zero. Would Dr. Aoki consider such an assay to have acceptable reproducibility since the estimated SD's are all close to zero? I would not.

If assessing the individual components contributing to the overall variability is viewed as a critical matter, then you could carry out a nested ANOVA to examine quantitatively the relative effects of variability among labs and variability among runs within labs on the overall response (e.g., the logEC50).

I could find nothing in Dr. Aoki's presentation to suggest how his approach could be used in a real world setting to determine whether or not an assay had acceptable reproducibility. One exercise that would be of interest would be to take a real world example and assess whether or not the assay has acceptable reproducibility in the usual way (considering CV's, etc.), and then ask Dr. Aoki and his colleagues to take the same data and make a similar 'bottom line' judgment based on his more complex assessment of weighted means, extracting sources of variability, etc. I strongly suspect that the same conclusion will be reached after considerably more work.

As a general rule, if a new complex statistical procedure is proposed to replace a 'less rigorous' one, then it should be demonstrated empirically how the old method fails and the advantages of the new approach in terms of the goal of the study, which in this case is accessing whether or not the assay has acceptable reproducibility. Until this is done, and concrete examples can be presented demonstrating the superiority of this more complex data assessment process, I see no need to make major changes in what is currently done.

Regarding Appendix 2, I strongly agree with Dr. Aoki that it makes no sense to calculate a CV based on log-transformed data. Surely, no one is recommending this (are they?). If so, this should be abandoned, and I agree with Dr. Aoki that the measure of variability to use in this case is the SD, not the CV. I further agree with his assertion that 'In general, CV is a good measure of variation where SD of a variable increases (linearly) with the mean of the variable. '

Dr. Aoki then states that 'there seems to be no reason to believe that the SD increases with the mean'. It is unclear if he is referring to the SD associated with the log transformed data (in which case I agree with him) or the untransformed data (in which case I disagree).

For example, toxic compounds with very low EC50's may have three runs with estimated EC50 values of (e.g.) 0.01, 0.03, and 0.05, while a non-toxic compound may have EC50 values of 1000, 3000, and 5000, In such cases, the SD's of the EC50's are quite different, but the SD's of the log transformed data are identical. This is what generally happens in practice. Thus, in terms of the EC50 I would use CV; in terms of the logEC50 I would use SD. I suspect that Dr. Aoki would agree with this.

Joe Haseman
5-25-06

------ End of Forwarded Message

June 20, 2006
Yutaka Aoki, ASPH Fellow at USEPA
aoki.yutaka@epa.gov

I share with Dr. Haseman the view that our primary goal is to evaluate whether the overall variability of the parameter estimate of scientific/regulatory interest from the assay is acceptably low. In the case of the transcriptional activation studies, for example, we are interested in whether the overall variability of the logPC10 across laboratories is acceptably low. In addition to this goal, it is often useful to have the capacity to evaluate the contributions of various sources of variability. In such cases it makes sense to have an estimate of intrinsic between-unit variability, not only overall (total) between-unit variability. (Please note that in my presentation I used the term "true between-run (lab) variation" to refer to what I am calling "intrinsic between-run (lab) variation" in this document.) In general, the overall (total) variability consists of two components: intrinsic between variability and overall within variability. That is, the following relationships hold:[1]

Overall (total) within-lab variability =
Overall (total) between-run variability = intrinsic between-run variability
+ overall within-run variability

and

Overall (total) between-lab variability = intrinsic between-lab variability
+ overall within-lab variability

Please note that the term "between-run (lab) variability" appears on both sides of the equations with different descriptors ("overall" vs. "intrinsic"). Hence there are two alternative interpretations for the term "between-lab variability," which appears in various assay validation guidelines as a standard component to be estimated in interlaboratory studies. I took the between-lab variation to mean intrinsic, not overall, variability, and applied the general, widely-used procedure for its estimation (i.e., the DerSimonian Laird random effects model). However, I realized from Dr. Haseman's comments that the term "between-lab variability" could be taken to mean "overall between-lab variability". What Dr. Haseman calls the "traditional procedure" is the natural procedure that ensues from this interpretation. Using one of these interpretations results in preference for a particular kind of between-lab variability estimate, the "overall" or "intrinsic".

There are a few potential uses for the complementary pair of estimates of intrinsic between-unit variability and within-unit variability as opposed to a single estimate of overall (total) between-unit variability alone. For instance, the pair of variability estimates are useful at a pre-validation stage when one is trying to identify specific sources of variation as a target of variability reduction. High variability in radioactive count measurement, for example, would tend to increase within-run variation, not intrinsic between-run variation. Inappropriate preparation of a stock standard solution for each run, from which appropriate serial dilution can be made reliably, would result in increase in intrinsic between-run variation, not in within-run variation. For an instance of post-validation use of the complementary variability estimates, suppose the overall between-lab variability for an assay has been found to be unacceptably high under a specified design and we would like to know how much an increase in the number of runs (or, rarely, labs) might reduce the variability to the desired level. Only with the estimates of intrinsic between-lab variability and overall within-lab variability (which is a function of the number of runs), would easy calculation of the necessary number of runs be possible.

As an additional benefit, the proposed procedure gives rise to a good estimator of overall variability, which in certain circumstances performs considerably better than the counterpart for the traditional method: the latter underestimates overall variability when intrinsic between-unit variation is small compared to within-unit variability. This difference arises because the two procedures handle standard errors (SEs) of estimates differently: our proposed procedure takes SEs of estimates (either run-specific summaries or lab-specific summaries that are to be further summarized) into account while the traditional method ignores them. The advantage of the proposed procedure was clearly noted in simulations I performed. In the case of the

[1] The relationships hold in terms of variance under the assumption of independence between the underlying components for the two right-hand side terms.

transcriptional activation data, for example, the overall between-laboratory variability would be more accurately estimated by the new procedure if the variability within each lab were large relative to the variability between labs. Underestimation of the overall variability is problematic since it gives a false sense of reproducibility to the user.

When deriving estimates of overall variability, which both Dr. Haseman and I regard as the most relevant variability measures, I obtain an estimate of intrinsic between-variability, and then combine it with within-variability estimates. This is done by taking into account the experimental design (i.e., how many runs and laboratories are actually used).

Although the new procedure may be more difficult to grasp conceptually than the traditional method of estimating overall variability, it is quite simple to implement. We consider the computational cost associated with our proposed procedure small, and particularly so when compared to potential benefits we gain by using it.

It is likely this response lacks the level of details that some readers would desire. I omitted many details for the sake of simplicity, but I am happy to provide more detailed information or answer questions upon request.

Appendix 4

Table 4.1 Summary of criteria that were not met according to ICCVAM Minimum Standard Procedures (ICCVAM 2003)

Minimum Standard Procedure	Met/ Not Met	Explanation and Justification
The stability of the test substances should be demonstrated prior to testing. In the absence of stability information, the stock solution should be prepared fresh prior to use.	NOT MET but resolvable retrospectively	The stabilities of test substances were not confirmed, however empirically stable substances were used. The stock solution was not freshly prepared. Under the inter-laboratory validation, the stock solution was prepared at the lead laboratory and then distributed to the participating laboratories. All stock solutions were stored at -20°C at each laboratory. The capabilities of the participating laboratories to make up stock solutions accurately were assumed, and the lead laboratory did not consider it necessary to include this as part of the validation process at the time. Should it be absolutely necessary for the purposes of the independent peer review, the participating laboratories could be requested to make up the stock solutions individually and then be subsequently assessed.
Studies should be performed in compliance with GLP guidelines.	NOT FULLY MET	The pre-validation was not to GLP, the inter-laboratory validation was under GLP, and the data collection for comparison with the ICCVAM list and hERa binding assay was not to GLP standards.
In a validation study, repeat studies would be conducted to evaluate intra-laboratory repeatability and reproducibility. In contrast, in screening studies, repeat studies are not conducted, except to clarify equivocal results.	NOT FULLY MET	The pre-validation and inter-laboratory validation was repeated but the data collection for comparison with ICCVAM list or hERa binding assay was not always repeated.

It should be noted that major deviation from the ICCVAM and ECVAM validation requirements could mean that the assay may not be considered by these validation bodies as correctly and formally validated for regulatory use.

Comments received from Drs Bill Stokes and Ray Tice (NICEATM) on Studies Conducted by CERI to Support the Validation of the hER-HeLa-9903 Estrogen Receptor (ER) Transcriptional Activation (TA) Test Method

Our comments are based on information CERI has provided in their report entitled, "Draft Pre-Validation and Inter-Laboratory Validation Report of the Human Estrogen Receptor Mediated Reporter Gene Assay", and other supporting materials, including those used to present information that CERI has provided at the request of the OECD Preliminary Validation Assessment Panel. Our assessment of the provided information is based on relevant information provided in Section VII of OECD Guidance Document No. 34, which recommends and defines the components of a new test method submission. Our assessment of the hER-HeLa-9903 ER TA test method protocol is based on the minimum procedural standards (we now call these essential test method components) recommended by ICCVAM[2] and based on the deliberations of an ICCVAM international expert panel on ER and androgen receptor binding and TA assays that met in May of 2002. Our evaluation of the substances used to evaluate the accuracy and reliability of the hER-HeLa-9903 ER TA test method is based on the ICCVAM list of recommended reference substances for ER binding or TA test methods[3].

Our comments are organized under the major headings in Section VII of OECD Guidance Document No. 34 as follows:

Introduction and Rationale for the Proposed Test Method

Reports and supporting materials address the rationale for the CERI ER TA test method, as specified in this section of the Guidance Document, but discussions regarding the specific limitations of the test method could be usefully expanded.

Test Method Protocol Components

A test method protocol has been provided, as specified in this section of the Guidance Document, but this is the protocol that was used for the experiments that involved multiple laboratories only. It is stated in the text that the in-house protocol was similar but the protocol followed throughout and any modifications and the rationale for those modifications needs to be included. For example, in the interlaboratory study, estradiol was tested over multiple concentrations but in the in-house studies, it was tested at only a single concentration. The rational for this difference should be provided.

In addition, in terms of the test method protocol, the highest concentration of substance tested was 10 µM, not the 1 mM recommended by the ICCVAM international expert panel and ICCVAM (see footnote 1). We appreciate that not all substances can be tested up to this concentration (due to solubility or excessive

[2] "ICCVAM Evaluation of *In Vitro* Test Methods For Detecting Potential Endocrine Disruptors: Estrogen Receptor and Androgen Receptor Binding and Transcriptional Activation Assays" (available at http://iccvam.niehs.nih.gov/methods/endocrine.htm).

[3] "ICCVAM Evaluation of *In Vitro* Test Methods For Detecting Potential Endocrine Disruptors: Estrogen Receptor and Androgen Receptor Binding and Transcriptional Activation Assays" and the 2006 Addendum to this report (available at http://iccvam.niehs.nih.gov/methods/endocrine.htm).

cytotoxicity) but the purpose for using this limit dose is to detect even very weak ER agonists or antagonists. Thus, at least some of the substances classified as negative by CERI have not been adequately tested (this was demonstrated in the data set provided by CERI for the last conference call) while others may have been adequately tested if solubility or cytotoxicity data can be provided to support the highest concentration tested.

There seems to be a lack of information in regard to the rationale/justification, criteria for use, and reliability for the cytotoxicity evaluation, which were conducted using the same basal cell line but with a different plasmid construct as a separate experiment. From verbal discussions, it appears that CERI does not feel a cytotoxicity evaluation is needed for the agonist tests. This issue needs to be formally discussed in their submission.

For use as a screening assay for ER or AR activity, it is critical that a TA test method evaluate for antagonist as well as agonist activity. Except for the intralaboratory repeat testing of three substances, an evaluation of the ability of the CERI ER TA test method to identify ER antagonists has not been provided. Furthermore, the antagonist protocol used in the testing of these three substances had no concurrent positive control, and did not use a reference standard with a full dose response curve as is done in the CERI agonist protocol. We appreciate the desire to move ahead with the agonist version of the test method independent of the antagonist version but wish to point out that a negative ER agonist study is virtually worthless without knowing whether or not the test substance binds to the ER and/or demonstrates antagonist activity. We do not agree with CERI's premise, stated in the most recent OECD teleconference, that the antagonist protocol is similar enough to the agonist protocol to be considered as validated in the same manner. We urge that the current ER antagonist protocol be modified to include appropriate positive controls and that further validation studies using this protocol be completed before peer review.

The protocol needs to include a discussion about potential "edging effects", and how to identify if the outside wells on the 96-well plate can be used because such effects are not detected under the experimental conditions used by a specific laboratory.

Characterisation and Selection of Substances Used for Validation of the Proposed Test Method

To facilitate validation of ER TA assay, ICCVAM compiled a list of 78 recommended reference substances. ICCVAM recommends that these substances be tested in a phased manner, with a minimum of 53 substances being tested across at least three laboratories. The remaining 25 substances are recommended for testing once in one laboratory or divided among two or more laboratories.

Our evaluation of the data submitted indicates that CERI tested a total of 56 substances, although only 10 were tested across multiple laboratories. Seven of these 10 substances are on the ICCVAM list and the remaining three have similar ER activities to other ICCVAM substances recommended for interlaboratory testing and could be considered as replacements for these.

Therefore, to meet ICCVAM recommendations, 43 additional substances from the ICCVAM recommended list or their equivalents would require further interlaboratory testing.

CERI tested 12 of the remaining 25 substances on the ICCVAM list that do not require interlaboratory testing at least once, leaving an additional 13 substances from the list or their equivalents that would require further testing.

Also, substances are not classified according to product class and only the 10 substances tested across multiple laboratories are classified by chemical class. These 10 substances represent 6 chemical classes compared to the 15 chemical classes represented by those substances recommended for interlaboratory testing by ICVAM (a total of 22 chemical classes are represented by the ICCVAM recommended list of 78).

In Vivo Reference Data Used to Assess the Accuracy of the Proposed Test Method

The comparison of experimentally derived results from ER TA agonist and immature rat uterotrophic studies conducted at CERI using 50 substances adequately supports the accuracy of the proposed ER TA agonist test method.

Testing all 78 reference substances would not only allow for a better characterization of the reliability and comparative sensitivity of the CERI test method versus other Tier 1 assays but also increase the likelihood that *in vitro* tests might be developed that could be used to reduce animal use in endocrine disruptor (ED) testing.

Test Method Data and Results

Results and data from prevalidation and interlaboratory studies conducted by CERI to support the validation of their hER-HeLa-9903 ER TA agonist assay have been provided, but much of this was not provided in the CERI draft validation report but rather at the request from the OECD preliminary validation assessment panel. It is assumed that the requested results and data will be included as appropriate in the appendices of the final validation report from CERI.

Test Method Relevance (Accuracy)

Because this test method is to be used as a Tier 1 screening assay (at least in the United States), there is no need for an evaluation of the ability of the test method to predict in *vivo* endocrine disruptor effects. However, such data are welcome and would allow better characterization of the ability of *in vitro* test methods such as this to reduce animal use in ED testing. The comparison of CERI derived ER TA results with ICCVAM published ER TA results for 46 substances is appropriate.

Test Method Reliability (Repeatability/Reproducibility)

In terms of intra- and inter-laboratory reproducibility, 10 substances (two strongly active positives, four moderately active positives, one weakly active positive, and three negatives) were tested three times in each of three laboratories. All tests were conducted using stock solutions provided by CERI (i.e., the full test method protocol was not evaluated). Furthermore, substances that posed potential problems in testing due to their physico-chemical characteristics (i.e., poor solubility) or because they were overtly cytotoxic were not tested. Thus, this is not an adequate evaluation of the intra- or inter-interlaboratory reproducibility of this test method. In its international evaluation of another ER TA test method, NICEATM/ICCVAM is proposing 12 substances to evaluate intralaboratory reproducibility in three labs (testing 3 times in each lab) and another 41 substances to be tested once in each of three labs to adequately evaluate interlaboratory reproducibility. These substances cover the range of anticipated agonist and antagonist responses, include a wide variety of chemical classes, and include substances with varied physico-chemical properties and cytotoxicity properties.

Also, in their interlaboratory evaluation, the reference substance, estradiol, was tested over its complete concentration response range. In contrast, for other substances, CERI tested estradiol at a single concentration. The former is recommended by the ICCVAM International ED Expert Panel and by ICCVAM for all experiments.

Test Method Data Quality

Interlaboratory studies testing 10 substances were conducted using GLP guidelines, but none of the pre-validation studies were conducted in this manner. At the last OECD preliminary validation assessment panel teleconference, CERI representatives indicated that a data audit has been recently conducted on the prevalidation studies and stated that non-compliance with GLP guidelines had no impact on data quality. We recommend that a specific discussion regarding data quality and non-compliance be included in the CERI report.

Animal Welfare Considerations (Refinement, Reduction and Replacement)

Our evaluation of the validation report and supporting materials indicate that specific discussions on how the proposed test method will refine, reduce, or replace animal use if used in a battery of tests to detect potential endocrine disruptors were not provided.

Practical Considerations

We recommend the inclusion of considerations such as the cost and time required to conduct the assay and report results. Considering the concerns about "edging effects", we also recommend expanding the discussion of necessary equipment and supplies, and the required level of training, expertise and demonstrated proficiency needed by study personnel.

Late Comments received on 3 June 2006 from Prof. Combes (member of the panel, but did not participate in the teleconferences or discussions prior to 3 June 2006).

Dear All,

Thanks for all the summaries which I have now had a chance to read in some detail, although I am afraid that I still have not had the opportunity to look at all the raw data.

My impression is that there has been an awful lot of work done on this assay and those involved deserve congratulations for their efforts and for getting us to the stage we are at.

Having said that, I have several overall concerns about the readiness of the work that has been done for peer review, since I am unsure as to the ability of the interlaboratory validation study to transparently and unequivocally demonstrate reliability and relevance of the assay for its stated purpose. In this regard, I share many of the concerns that have been raised in the NICEATM comments raised during the last teleconference as presented in Appendix 1 of the latest set of minutes.

Due to the large amount of information and data, I am unclear as to exactly where we are now and welcome the suggestion that there should be an overall report. This could well serve as the document for eventual peer review, but this decision should not be taken until we have all seen the document and agreed on its status. The last thing we would want is for the peer review report to be controversial (as indeed is the report for the Uterotrophic assay) as this would undermine the validation process and give the assay a bad name, when it could all be avoided by being less hasty and ensuring that the validation study is as good as possible.

I personally remain unconvinced that the stuies are ready yet for peer review for the following main reasons:

1. the raw data are not as transparent as they should be
2. there is a need to agree on how the data are transformed and statistically analyzed (personally I prefer the presentation of straightforward error bars)
3. it appears to me that the validation has only been performed in Japan, when for it should be assessed in other countries (this is no criticism of Japanese laboratories, merely it is necessary to ensure that reliability extends to other countries
4. there have been claims for the deviation of the studies from accepted OECD, ECVAM and ICCVAM validation criteria - these need to be discussed in more detail.

With regard to other matters, I think that it would be good to have more detail concerning what was discussed in relation to the assay at the recent WNT meeting. In addition, I am unhappy with the vagueness of what is stated regarding the potential arrangements for peer reviewing the assay, as stated in the minutes of the last teleconference.

A peer review of a validation study should not be contracted out to a laboratory, for goodness sake!

I am also very concerned that the OECD might be asked to organise a peer review, in view of the debacle over the review of the uterotrophic assay. Peer review of new in vitro methods should be left to those with experience and authority with undertaking them in conjunction with relevant legislative authorities; namely ICCVAM and the ECVAM Scientific Advisory Committee. In fact, my suggestion would be for a joint peer review organised by ECVAM, ICCVAM and the newly-formed JACVAM. This would be an excellent opportunity to initiate a world-wide peer review study and to capitalise on the existence of these centres. However, I re-iterate that no peer review should be undertaken until it can be
ensured that the validation study meets all the necessary criteria.

I apologise if I seem rather over-critical, but I am not trying to be - I am very impressed by the work achieved on the assay, but I think we should be cautious in going too fast and losing the opportunity to build on the excellent foundation that we have. I am as keen as anyone to see these types of assays on the books to augment and eventually replace the in vivo methods. But we must get it right, ensure it meets international criteria, and check that everything is independent and transparent.

I hope all this helps, with best wishes,

Bob Combes

Detection of anti-estrogenic activity using reporter gene assay

Description: This document provides a methodology for detecting anti-estrogenic activity of chemicals by reporter gene assay technique using hER-HeLa-9903 cell line.

Materials and methods

1. Test chemicals

Test chemicals should be dissolved in in dimethylsulfoxide (DMSO) at a concentration of 10 mM.

2. Competitive substance

17β-Estradiol (E2)

3. Vehicle for chemical stock solutions

Dimethylsulfoxide (DMSO) should be used for the vehicle.

4. Test system and operating procedures

4.1 Cell lines

hERα-*HeLa*-9903 stable cell line (Sumitomo Chemicals Co.) will be used for the assay and 9903-control cell which consistently express firefly luciferase by the RSV promoter without stimulation will be used for evaluating cell-toxic effect of chemicals when anti-estrogenic like effect is observed.

4.2 Cell culture (See support protocols No.1 – No. 4)

Cells should be maintained in Eagle's Minimum Essential Medium (EMEM) without phenol red, supplemented with 10% dextran-coated-charcoal (DCC)-treated fetal bovine serum (DCC-FBS), in a CO_2 incubator (5% CO_2) at 37°C.

4.3 Preparation of chemicals

All chemicals will be dissolved in DMSO at a concentration of 10 mM, and the solutions will be serially diluted with the same solvent at a common ratio of 1:10 to prepare stock solutions with concentrations of 1 mM, 100 μM, 10 μM, 1 μM, 100 nM and 10 nM. In the case of positive control substance (E2), stock solutions will be prepared at concentrations of 100 μM, 10 μM, 1 μM, 100 nM, 10 nM, 1 nM and 100 pM.