OPG/BP 2010 EVS METHODOLOGY FOR CALCULATION OF NOP TRIP SETPOINT:

INDEPENDENT VERIFICATION AND BENCHMARKING OF STATISTICAL METHOD AND MATHEMATICAL FRAMEWORK

Anthony O'Hagan

13 March, 2013

EXECUTIVE SUMMARY

EVS 2010 is a statistical tolerance limit solution to the NOP trip setpoint problem. This report presents my evaluation of the theoretical correctness and practical fitness for purpose of EVS 2010 for that problem. On the basis of a technical review of the EVS 2010 theory and three separate benchmarking exercises to assess its fitness for purpose, this report reaches a number of conclusions supported by nine principal findings.

> *My technical review found that the EVS 2010 theory is mathematically and statistically correct, and the benchmarking found that in the great majority of tests EVS 2010 provided adequate protection in terms of the tolerance limit coverage.*

In the context of the NOP trip setpoint problem, 'adequate protection' means that the method provides at least 95% assurance of a trip operating early enough during a slow loss of regulation event to prevent the risk of dry-out in 95% of those events. These positive findings suggest that EVS 2010 is basically sound, at least in theory, and has the potential to provide a practical solution to the NOP trip setpoint problem.

> *However, a number of other findings make it clear that at present there are several outstanding concerns regarding its use for determining NOP trip setpoints in practice.*

In particular, benchmarking indicates that there are three specific situations where EVS may not provide adequate protection against the risk of dry-out. Furthermore, the inherent limitations of the generic benchmarking employed in this work mean that the encouraging performance in benchmarking tests may not hold for a particular real application.

> *In any proposed practical application it will be essential to provide assurance that EVS 2010 will deliver good tolerance limit behaviour in that application.*

My "verification and benchmarking" task formally yields only these conclusions, and in particular is not prescriptive about how the assurance demanded in this third conclusion should be provided. However, the section in my report entitled "Summary, Conclusions and Recommendations" ends with some recommendations for future progress in the NOP trip setpoint problem, including three ways to achieve assurance of good tolerance limit behaviour in proposed applications of EVS 2010.

CONTENTS

INTRODUCTION

BACKGROUND

This document constitutes my *Final Report* to the Canadian Nuclear Safety Commission (CNSC) under Contract 87055-10-1226 – R396.2: "Independent Verification and Benchmarking of Statistical Method and Mathematical Framework in OPG/BP 2010 EVS Methodology for Calculation of NOP Trip Setpoint".

The EVS methodology is set out in the document "A Genuine '95/95' Criterion for Computing NOP Trip Set-points Using EVS Methodology" by Paul Sermer and Fred Hoppe. That document is report number G0263/RP/008 from AMEC NSS Ltd., dated September 30, 2010, and will be referred to herein as the EVS Report. The methodology that it proposes will be referred to as EVS 2010. The EVS Report was revised in 2011 in AMEC NSS document number G0263/RP/008 R01, dated October 4, 2011. The basic method remains as originally presented in the 2010 document, but with some refinements to the technique of implementation. Where it is helpful to comment on those changes, this document will be referred to as the EVS 2011 Report.

The contract identifies three elements to the "verification and benchmarking" of EVS 2010. The first is a technical evaluation of the correctness of the mathematical theory of EVS 2010. The second and third elements are two groups of benchmarking tests to evaluate the performance of EVS 2010 in practice. In the course of this contract, I have prepared separate reports on each of these exercises, which are attached as appendices to this Final Report. The body of this report is a narrative that explains all the key findings from those exercises and draws final conclusions, but the reader may find further details in the appendices.

THE NOP TRIP SETPOINT PROBLEM

The NOP (Neutron Overpower) trip is designed to protect a Candu reactor core from dry-out in conditions of slow loss of regulation (LOR). A very brief overview of the NOP trip setpoint problem is given here, with details set out in the subsequent section, "Terminology and Notation".

The reactor core comprises a large number of fuel channels and the reactor state at any one time includes the power being produced in each channel. When a channel is producing power in excess of a value known as the critical channel power (CCP) for that channel, then dry-out may occur. Consequently, the objective of the NOP trip is to prevent the channel power in any channel from exceeding its CCP value.

In a slow LOR event, the core heats up relatively slowly and evenly. The NOP trip is an intervention that is automatically applied to shut the reactor down when an appropriate flux detection measure (FDM) based on flux detector readings from within the core reaches a certain value. That value is known as the installed NOP trip setpoint (TSP). The term 'installed' refers to the fact that this value is fixed in the operating software, in order for the trip to operate automatically when required.

Ongoing processes of refuelling and fuel depletion mean that in practice the detailed state of the reactor core is continually varying. When this is added to the fact that there are inevitably errors in contemporaneous measures of power and in flux detector calibration, the precise condition of the reactor when the trip is activated (by the FDM reaching the installed TSP) will be different from time to time. In particular, it may trip very early, when the powers in all channels are still well below their CCP values, or it may trip too late, when at least one channel power has already exceeded the critical value.

This variability in operating conditions from time to time is inevitable and means that choosing a suitable value for the installed TSP is a statistical problem. EVS 2010 is a statistical method that is intended to derive a suitable TSP value.

TERMINOLOGY AND NOTATION

MARGIN TO DRY-OUT

At any given moment in the operation of the reactor, we can define the critical power ratio (CPR) for each fuel channel as the ratio of the CCP to the current power level of that channel. If the CPR for any channel is less than 1, then the channel power exceeds the critical value CCP and there is a risk of dry-out. The margin to dry-out (MTD) is defined as the minimum of the CPR values for all the reactor fuel channels. The objective of the NOP trip is to ensure that the MTD at trip is greater than 1.

FLUX DETECTION MEASURE

The MTD cannot be observed in practice, and the trip instead operates on the basis of the readings from flux detectors in the core at any given moment. The readings from the various flux detectors are reduced to a single value that we will call the flux detection measure (FDM), and the trip operates if the FDM is greater than or equal to the installed TSP.

The formula for the FDM may depend on the reactor, and in particular on how many flux detectors are installed and their arrangement into safety channels. According to a typical formula the FDM is the minimum over the safety channels of the maximum flux detector reading from any of the detectors in a given channel. The effect of this particular FDM is that the trip operates if in every safety channel there is at least one flux detector reading that reaches or exceeds the installed TSP.

It should be noted that in practice flux detector measurements are subject to errors of measurement and/or calibration, so the value of the FDM that is computed from actual detector readings will differ from the true FDM value that would arise from applying the formula to the true flux/power at the flux detector locations.

THE IDEAL TSP

In principle, if we knew everything about the reactor conditions at a given time point we would know the current true value of the MTD and also the current true value of the FDM. Then we could deduce an exact TSP that is just the product of the current true values of MTD and FDM. The reasoning for this is simply that in a slow LOR event the power is assumed to rise uniformly throughout the reactor. If overall power rises by a factor equal to the current true MTD, then the margin to dry-out will reduce to 1, which is the point at which the NOP trip needs to operate. The same uniform rise in power would raise the FDM by the same factor, and so the FDM would increase to the product of its current value with the current MTD.

We will refer to the resulting TSP as the ideal TSP. Some features of the ideal TSP should be noted.

1. The ideal TSP is not a constant. It depends on current conditions, particularly on ripples and flux shape (which are discussed more fully in a later subsection), and these are unknown at the time of trip operation. The installed TSP has to be a constant, so in practice it is inevitable that in any given instance it will be either lower than the current ideal TSP (and so cause the trip to operate too early) or higher than the ideal TSP (in which case the trip will operate too late, after the MTD has dropped below 1).
2. The NOP trip operates on the measured FDM, which is not the true value because of errors in detector readings. If the measured value is below the true FDM the trip will operate later, and conversely if the measured value is above the true FDM the trip will operate earlier.

A-UNCERTAINTIES

The NOP trip will operate whenever the measured FDM reaches the installed TSP. At any given instance, the combination of the detailed reactor conditions and the flux detector errors may lead to the trip operating too early or too late. The following are two equivalent definitions of a trip operating too early:

- The ideal TSP-at-trip is higher than the installed TSP.
- The MTD-at-trip is higher than 1.

Imagine that it would be possible to operate the reactor for an arbitrarily long period of time (without any change or deterioration in operating conditions, and with a fixed installed TSP) and observe an extensive number of trip events. Then the collection of values of MTD-at-trip would form a probability distribution, such that any single event could be considered to be a random draw from this distribution. The proportion of such events in which the MTD-at-trip is less than 1 would measure the probability that the installed TSP fails to protect against the risk of dry-out that arises when MTD goes below 1.

It is important to understand the nature of this probability distribution. At any given NOP trip event, we would be uncertain about the true MTD-at-trip which has occurred. (We would no doubt learn whether the trip has operated in time, but still the actual MTD value would be unknown.) Statisticians characterise uncertainty using probabilities, and the probability distribution of MTD-at-trip is a formal description of this uncertainty for a future actual NOP trip event. It does not necessarily mean that MTD-at-trip is a random variable according to the conventional notion of randomly occurring phenomena, although it may do. It is enough that the MTD-at-trip for a future actual event is uncertain and that we describe this uncertainty with a probability distribution.

Because the uncertainty about MTD-at-trip is uncertainty about a future *actual* event, it will be referred to here as A-uncertainty. In the EVS report, the authors call it *aleatory* uncertainty. To use such a term introduces philosophical questions that are unhelpful and distracting, so I prefer not to use it. My choice of the term A-uncertainty has the mnemonic value of relating to the NOP report's terminology, but it is most helpful to remember that A-uncertainties relate to the *Actuality* of a future NOP trip event.

We can equivalently characterise A-uncertainty through a distribution for the ideal TSP-at-trip. It would be a different probability distribution from the A-distribution of the MTD-at-trip, but because of the equivalence noted above the probability that the ideal TSP-at-trip is less than the installed TSP is an equivalent way of calculating the probability that the installed TSP fails to protect against the risk of dry-out. In practice, the fact that the A-distribution of the ideal TSP-at-trip does not depend on the installed TSP makes it slightly easier to work with when attempting to set a suitable installed TSP than working with the A-distribution of the MTD-at-trip.

COMPONENTS OF A-UNCERTAINTY

A-uncertainty arises from uncertainty about the detailed configuration of channel powers and from errors in flux detector readings or calibration. The configuration of channel powers in the reactor core is conventionally understood as comprising three factors. One is the nominal channel powers, which represents the overall mean pattern over time, and for instance the nominal channel powers are higher in the centre of the core than at the periphery. The second factor is termed the ripples, which are considered local perturbations of the nominal pattern, arising primarily from the history of refuelling actions. The third factor is termed the flux shape and represents potentially more substantial and less local skewing of channel powers, arising from other control actions (and possibly inactions).

Whereas the nominal channel powers are known, the ripples and flux shape will vary over time and are unknown at any given time. These unknown factors also affect the critical channel powers and true flux detector values.

It is helpful to consider the sources of A-uncertainty in three components.

1. Ripples, denoted by Q.
2. Flux shape, denoted by Φ.
3. Flux detector errors and other uncertainty elements as appropriate, denoted jointly by Θ.

Uncertainty about ripples, flux shape and other components is characterised by A-distributions for Q, Φ and Θ. The ideal TSP can be formally written as a function f(Q, Φ, Θ) of these components, and its A-distribution is mathematically implied by this function and the A-distributions of Q, Φ and Θ. The function f is rather complex, as set out in the EVS 2010 report, and involves operations of maximising and minimising (of critical power ratios and flux detector readings). Nevertheless, in principal, if we knew the A-distributions of Q, Φ and Θ we could deduce the A-distribution of the ideal TSP.

THE REFERENCE TSP

The choice of installed TSP is a decision that must be taken in recognition of the consequences of a given choice. Setting a very high installed TSP will result in too large a risk of the trip not operating in time to prevent dry-out. Setting a very low installed TSP will give only a small risk of the trip not operating in time, but instead will result in an economic cost associated with derating in order to operate the reactor within tighter constraints. It is the A-distribution of the ideal TSP that provides the important quantification of the risk of not tripping until MTD has fallen below 1.

In practice, safety is a prime consideration. The objective is always to make the A-probability of the installed TSP being below the ideal TSP, which we denote by γ, large. Typical values are $\gamma = 0.95$ or $\gamma = 0.99$. From the desired value of γ and the A-distribution of the ideal TSP, the reference TSP is determined – it should be the value such that there is an A-probability of γ to the right of the reference value. The reference TSP for a given γ is denoted by t_γ.

FLUX SHAPE UNCERTAINTY

The A-distributions define the context within which the performance of the installed TSP is to be controlled by the γ parameter. If, say, $\gamma = 0.95$, then the requirement is that we wish the installed TSP to be set so that on at least 95% of occasions it trips early enough. The A-distributions define the meaning of that 95%, they answer the question, "95% of *what* occasions?"

Because flux detector errors and ripples vary from time to time the values of Q and Θ that pertain in a future slow LOR event can be treated as random. The A-distribution of Θ is defined by the distribution of random flux detector observation and calibration errors. The A-distribution of Q is defined by the variation of ripples from time to time. The nature of the A-distribution of flux shapes is less obvious.

The treatment of flux shapes as random is somewhat controversial. From one perspective, it is certainly the case that in an actual future slow LOR event the flux shape pertaining at the time is uncertain. However, flux shape is not entirely unpredictable, and one approach is to group the possible flux shapes into classes, with the understanding that at any time it may be known which class of flux shape holds. Then different installed TSPs can be set for different flux shape classes. Nevertheless, there remains the question of whether flux shape is random within a class, and if so how the within-class probabilities or weights may be determined. I have identified three possible ways of answering this, each corresponding to a different answer to the question, "95% of *what* occasions?"

1. Flux shape is treated as genuinely random (within a class), and the weights are the proportions of slow LOR events on which any given flux shape would hold. The tolerance limit is designed to produce a trip early enough to avoid the risk of dry-out on 100γ% of events in which Q, Φ and Θ are all varying according to the specified A-distributions. The weights need to be estimated empirically.
2. Flux shape is not random. The installed TSP should be set at the minimum value that the tolerance limit would give for any flux shape (within a class). The tolerance limit is designed to produce a trip early enough to avoid the risk of dry-out on *at least* 100γ% of events in which Q and Θ are all varying according to the specified A-distributions, for all possible Φ (within the class). There are no weights to estimate.
3. Flux shape is not random. The tolerance limit is designed to produce a trip early enough to avoid the risk of dry-out on *an average of* 100γ% of events in which Q and Θ are all varying according to the specified A-distributions, averaged over all possible Φ (within a class). This is equivalent to treating flux shape as random but with equal probabilities (within the class). The weights are fixed and do not need to be estimated.

In case 3, it is conceivable that unequal weights would be specified, so that the average referred to becomes a weighted average rather than a simple average.

The EVS document assumes that the objective is to address case 1 or 3, i.e. that Φ is to be treated as random (with weights which may or may not need to be estimated). However, case 2 could be addressed using a simpler version of the EVS theory that is presented in Section 5.3 of the EVS document.

E-UNCERTAINTY

Unfortunately, in practice the A-distributions are not known, and so t_γ is also unknown. We therefore have to recognise another area of uncertainty in the problem. This is uncertainty about the A-distributions due to lack of information. In the EVS report this is referred to as *epistemic* uncertainty, but again I prefer to avoid the philosophical content of such a term and instead refer to it as E-uncertainty. I believe it is helpful to think of E-uncertainty as arising from the imperfection in the available *Evidence* regarding the A-distributions.

In principle we have E-uncertainty concerning the A-distribution for each of the components Q, Φ and Θ. It is important to recognise, though, that there are quite different levels of E-uncertainty in each case.

A-distribution of Θ. The nature of errors in flux detector readings is relatively well known. EVS 2010 assumes that the A-distribution of Θ is completely known, and therefore EVS 2010 does not admit E-uncertainty about this component.

A-distribution of Φ. Each flux shape is associated with a particular operating condition. It is generally assumed that the set of possible flux shapes for a particular reactor can be enumerated exhaustively, although in practice a smaller set of representative flux shapes may be used. By the application of a suitable physics code, an estimate of the flux shape associated with every one of the possible (or representative) operating conditions is available. Because of computational inaccuracies, there is uncertainty about each possible true flux shape. The E-distribution for these computational inaccuracies is partially understood, and in EVS 2010 this E-distribution is assumed known. Note that this E-uncertainty regarding the possible true flux shapes applies in all three cases discussed above, including cases 2 and 3 where Φ is not formally regarded as random.

In cases 1 and 3 the A-distribution of Φ comprises both the set of possible flux shapes and the probabilities or weights associated those shapes. In case 1 the weights are in reality unknown, so this should be another source of E-uncertainty, but these weights are assumed known in EVS 2010.

A-distribution of Q. The ripples vector Q has a value for each fuel channel, and the possible configurations of ripples are in practice almost limitless. Available evidence consists of a number of "observed" ripples vectors (denoted in the EVS Report by S), assumed to be a random sample from the A-distribution of Q. However, these "observations" are again computed by a suitable physics code, and so the "observed" S vectors are subject to computational inaccuracies. E-uncertainty about the A-distribution of Q therefore arises both from the fact that we only have a sample of instances from that distribution (which is otherwise completely unknown) and from the fact that each instance is "observed" with computational inaccuracies. In EVS 2010 the E-distribution of the computational inaccuracies is assumed known.

TOLERANCE LIMITS

The reference TSP t_γ is a function of the A-distributions of Q, Φ and Θ, but there is E-uncertainty about those A-distributions. An E-distribution for t_γ can in principle be derived from those E-uncertainties, and this becomes the statistical basis for computing a suitable installed TSP.

The computed TSP, denoted in the EVS Report by W, is chosen such that there is an E-probability of β that the chosen value is lower than the true t_γ. Again, safety concerns dictate that β should take a high value, typically $\beta = 0.95$ or $\beta = 0.99$.

This is the definition of a statistical tolerance limit.

It may be helpful to summarise the ideas so far, in order particularly to clarify the roles of the two different forms of uncertainty.

- The *Actual* conditions that will pertain at any given point of time in the future operation of the reactor are uncertain. In particular, it is uncertain what will be the ideal TSP at that *Actual* time, and this uncertainty is A-uncertainty. It is intrinsic and irreducible uncertainty induced by the fact that conditions vary from time to time and so the ideal TSP also varies from time to time.
- Because of A-uncertainty, we cannot set an installed TSP such that it will always cause a trip just in time to prevent risk of dry-out. Instead, the objective is the reference TSP value t_γ, which is such that there is a (suitably high) A-probability γ that the ideal TSP will be higher than t_γ. This ensures that there is a (high) A-probability of the trip being triggered in time to prevent risk of dry-out.
- The sources of A-uncertainty are ripples Q, flux shape Φ and flux detector errors (and possibly other factors) Θ. If the A-distributions of these components were known we would be able to determine the reference TSP and set this as the installed TSP.
- However, those A-distributions are not completely known. In particular, we have estimates of the possible flux shapes but these are "observed" with computational inaccuracies, and we have only a sample from the A-distribution of ripples and these are also "observed" with computational inaccuracies.
- Because the *Evidence* concerning ripples and flux shapes is imperfect, there is E-uncertainty, which is not irreducible. It may be reduced by obtaining more or better evidence.
- In particular, there is E-uncertainty about the true value of the reference TSP t_γ. As a result, statistical methods are used to obtain a computed TSP W, such that there is a (suitably high) E-probability β that W is lower than t_γ. This computed TSP is proposed as the appropriate method to set the installed TSP.

In short, the tolerance limit is a value W that with high probability or confidence (β) will be low enough to cause the NOP trip to operate in time in a high proportion (γ) of actual instances.

In my technical review I stated my view that a statistical tolerance limit is a sound approach to managing risks in a context such as that of the NOP trip setpoint problem. I believe that the separation between A- and E-uncertainties that is intrinsic in the tolerance limit approach is sound and makes the nature of the risks and the regulatory choices clear. The problem to be solved by EVS 2010 or any other statistical tolerance limit method is defined by γ, β and the nature of the A-uncertainties (particularly regarding Φ). In my review I emphasised the importance of a clear understanding in the industry of these concepts so that these defining characteristics of the problem can be chosen in a transparent and informed way. My impression is that this understanding is as yet only partial.

THE EVS SOLUTION

EVS 2010 is a statistical tolerance limit method. The mathematical theory in the EVS Report presents formulae with which W can in principle be calculated (based on available evidence and for any given values of β and γ. The formulae cannot be applied simply as given in that report because they require values for some quantities which will not in practice be known. So the application of the EVS 2010 method requires additional steps to estimate these quantities. And like any statistical method it is dependent on some modelling assumptions.

The contract title states that my task is to conduct an "Independent Verification and Benchmarking of Statistical Method and Mathematical Framework" of EVS 2010. The reference to both "statistical method" and "mathematical framework" mirrors the distinction between the mathematical theory of EVS 2010 and the methodology of applied statistics employed in putting the theory into practice. This Final Report addresses both of those aspects.

TECHNICAL REVIEW

The full text of my technical review is given in Appendix A to this report. Some of the concerns raised therein have been at least partially resolved in subsequent discussion, while others have since gained in prominence. These changes of emphasis are reflected in the summary given here.

EVS THEORY

The mathematical and statistical theory of EVS is set out in detail in the EVS Report. My initial lack of familiarity with Candu reactors and the NOP trip setpoint problem meant that I had many difficulties in developing a full understanding of the theory, but I found that people at CNSC, OPG, Bruce Power and AMEC NSS were unfailingly helpful with their explanations and patient with my misunderstandings. I take this opportunity to express my thanks to all those people.

A major finding of my review is that the theory is mathematically correct.

However, a decision as to whether EVS 2010 is a valid and suitable method to use in practice for the setting of a reactor's installed TSP must depend on more than this. All statistical methods are based upon an assumed model, and even if mathematically correct their performance in practice will depend on how well the assumed model fits the reality of the application. Reality is always more complex than any model – indeed that is the essence of the term 'model' – and so all statistical methods in practice make assumptions that are not strictly valid for the application. The practising statistician must make a judgement as to whether a proposed method is *fit for purpose*, that is, whether the assumptions are sufficiently close to reality for the method to deliver answers that are accurate enough for the application. In this sense, the practice of statistics is an art as much as a science.

It is also common for statistical methods to be approximate, in the sense that the theory itself makes approximations. For instance, a huge body of applied statistical analysis uses generalized linear models, for which all the theory is approximate. EVS 2010 also makes some approximations of this kind. Again, a judgement of whether EVS 2010 is fit for the purpose of determining NOP trip setpoints requires an assessment of whether the approximations are accurate enough in the context of the actual application.

My review therefore highlighted the assumptions, approximations and simplifications (together referred to as compromises) made by EVS 2010 and emphasised that their effect in practice would need to be assessed in order to decide on its fitness for purpose. The review recommended that the practical impact of several of these compromises should be evaluated with the aid of benchmarking tests.

I will enumerate here the various compromises that are part of the EVS theory. It is important to be aware that some of these are not specific to EVS 2010 but will have to be addressed by any statistical method attempting to solve the NOP trip setpoint problem.

ASSUMPTION: SUPERPOSITION AND INDEPENDENCE

EVS 2010 adopts a principle known as superposition which is widely accepted in the industry, albeit as an approximation to reality. The superposition principle asserts that the vector of channel powers at any given time can be decomposed into the product (channel by channel) of three vectors, the nominal channel powers (which are fixed and known), the ripples and the flux shape. Beyond this fairly innocuous assertion is the implication that the ripples and the flux shape have distinct and identifiable origins. That implication is further taken in the EVS theory to imply that the A-distributions of Q and Φ are independent.

This independence and the superposition principle underlying it should be noted as an assumption which in reality probably does not hold exactly. It may be that there is empirical evidence to support the principle, but that evidence would presumably come from normal operating conditions. Slow LOR events are unusual and I imagine are reacted to quickly enough for no useful data to be available to assess whether the principle is valid in such an event.

Because superposition is apparently an accepted assumption in the industry no tests were specified in the benchmarking exercises to examine, for instance, the effect of correlation between Q and Φ. If the assumption were to be called into question, then it would be necessary to examine the robustness of both EVS and all other currently considered methods for the NOP trip setpoint problem to departures from superposition and independence.

ASSUMPTION: PHYSICS CODE ERROR STRUCTURES

E-uncertainties are primarily driven by computational inaccuracies in the physics codes whose outputs are used as evidence from which to learn about the A-distributions of Q and Φ. In order to analyse those data correctly and derive the E-distribution of t_γ, it is necessary to make assumptions about the structure of errors in the computations. EVS assumes specific forms of E-distributions for physics code errors

Note that we use the word ‘error’ here in the usual statistical sense of the difference between a value computed by a physics code and the true value of the corresponding quantity. That difference may be due to the mathematical theory underlying the code being a wrong or incomplete representation of reality, to the inevitable imprecision of the algorithm used to solve the theoretical equations, to rounding errors in the computer, or to any other cause such as operator error. These errors are in reality usually deterministic, because running the code twice with exactly the same inputs will usually produce exactly the same outputs. However, it is normal to treat the errors as random. This in itself is perhaps another assumption which does not hold in practice. Nevertheless it is acknowledged to be a reasonable approximation in the sense that errors generated by large computer codes are unpredictable and for most purposes behave as if they are random (in the same way as pseudo-random numbers look random but are really deterministic).

The EVS theory effectively assumes that these E-distributions are known. Strictly, it assumes only that means and variances are known for induced errors in other computations, but as will be discussed in the section on *EVS Implementation* these are derived in practice from fully specified E-distributions for the original code errors. The benchmarking exercises included various tests to investigate EVS performance under varying specifications and mis-specifications of error structures.

ASSUMPTION: KNOWN FLUX SHAPE WEIGHTS

The EVS theory assumes that the number of possible flux shapes is known, that there is an estimate from physics codes for the effects on channel powers of each of the possible flux shapes, and that the weights/probabilities of the various flux shapes in the A-distribution of Φ are known.

In reality, the set of flux shapes for which estimates are available will usually be a representative subset of all possibilities. Some investigation of robustness to mis-specification of the set of flux shapes was undertaken in the benchmarking exercises. The assumption of known weights, or probabilities, is innocuous under case 3 discussed above, but is a strong one under case 1.

ASSUMPTION: FLUX DETECTOR ERRORS

The A-distribution of Θ, which arises from observation and calibration errors in the flux detector readings, and possibly from other sources, is assumed to be known. Although this assumption will also be wrong in practice, the magnitude of these errors is such that the A-uncertainty in the ideal TSP is dominated by the uncertainty in ripples and flux shapes. Accordingly, the benchmarking exercises did not consider robustness to mis-specification of this distribution.

ASSUMPTION: KNOWN VALUES OF ASSORTED CONSTANTS

The final formulae for the tolerance limit (W) in the EVS theory are expressed in terms of a number of constants whose values must be known in order to compute W. In the primary equation:

- r_γ is a constant linking the reference TSP tγ to the mean and variance of the A-distribution of the ideal TSP.
- K_3 and K_4 are third and fourth moments of the 'best estimate' V variable.
- κ is the ratio between the standard deviations of the ideal TSP and V.

A second equation expresses κ in terms of several other constants. In reality, none of the constants will be known and so will need to be estimated.

APPROXIMATION: CENTRAL LIMIT THEOREM

The EVS theory employs an approximation in the form of a multivariate central limit theorem. The approximation will be accurate if the number of sample ripples "observations" S is sufficiently large.

EVS IMPLEMENTATION

Any implementation of the EVS theory must address all of the assumptions in the theory, and where numerical values are required these must be sourced.

- The superimposition and independence assumption should be acknowledged and judged to be valid, or at least to be an acceptable approximation.
- Physics code error structures must be specified. The forms of distributions, usually either normal or lognormal, must be specified, together with relevant parameters such as means and variances. Of particular importance are covariances. In practice these are unknown. Distributional forms are assumed and parameters estimated on the basis of limited empirical evidence on the accuracy of the physics codes.
- The distribution of flux detector observation and calibration errors, as encoded in the variable Θ, must be specified. Again, this entails identifying distributional forms and relevant parameters such as means, variances and covariances. And again, these are in practice assumed or estimated on the basis of limited empirical evidence.
- The nature of the flux shape uncertainty must be specified. If Φ is to be treated as random, then either the weights must be pre-specified or else will need to be estimated. In the latter case, substantial uncertainty regarding the weights can be expected in the light of the very limited empirical evidence.
- If the assumed set of possible flux shapes is seen as representative of the possibilities but in reality incomplete then this should be acknowledged and judged to be an acceptable approximation.

- The various unknown constants required in the formulae for the tolerance limit W must be estimated. In practice this is done using a variety of approximations, including the so-called surrogate methodology.

It is clear that a relatively large number of quantities will in practice be estimated. These will only approximate to the true values which are assumed in the EVS theory. Because of these approximations and because of the approximation in the theory from using the multivariate central limit theorem, EVS 2010 in practice will not be an exact tolerance limit. Whereas the theory says that with E-probability β the computed TSP will be less than the reference TSP t_γ, in reality this probability will be more or less than β. The actual coverage will depend on how well the approximations work in any particular application.

It would be preferable to address the approximation arising from estimation of required quantities in a formal statistical way. Estimation errors increase the variability of the computed TSP, and if this additional uncertainty is not formally addressed statistically it will tend to result in the coverage being less than β (or in the language to be introduced in discussion of the benchmarking results, the non-coverage will exceed 1 – β). The EVS 2011 Report introduced an attempt to address statistically the magnitude of errors introduced by estimating some of the unknown quantities. The mathematics was complex and the adjustment a somewhat ad hoc downward shift of the computed TSP, but with the sound intention of counteracting the tendency for estimation errors to increase non-coverage.

These various considerations clearly imply that the performance of EVS in practice can be expected to differ from the tolerance limit properties given by the theory. There is of course always a difference between theory and practice, and no statistical method retains in practical application the nice properties that its theory says it should have. The question is always whether, in the specific applications for which it is proposed to use it, the method’s performance is close enough to what the theory says. If so, we would say that it is fit for purpose.

The benchmarking exercises were therefore a vital part of the assessment of the fitness of EVS 2010 for the purpose of setting NOP trip setpoints.

BENCHMARKING EXERCISES

PURPOSE AND PRINCIPLES

Benchmarking is generally understood to be a process of measuring performance against that of relevant competitors or comparators. EVS 2010 is a statistical tolerance limit method designed to be applied to the problem of computing NOP trip setpoints. In order to benchmark it we therefore need competitors or comparators against which it might reasonably be evaluated. One immediate difficulty, then, is the fact that there are essentially no other statistical tolerance limit methods designed to be applied to this problem.

The benchmarking exercises conducted under the present contract addressed this difficulty in a number of ways.

- In some tests a new comparator was devised specifically for the purpose of benchmarking EVS 2010. The disadvantage of this approach was that it could not be expected that a method quickly developed under this contract might be a serious competitor for EVS 2010, which has been developed using many man-months over a period of several years.
- In some tests EVS 2010 was compared with some simple variants of existing approaches for computing NOP trip setpoints. This is also somewhat unsatisfactory because these methods were not designed as proper statistical tolerance limits.
- In all tests EVS 2010 was compared with absolute standards of performance of statistical methods generally, and of tolerance limits in particular. This form of benchmarking, absolute rather than comparative, is also known as validation. In view of the limited value of the above comparators, the benchmarking of EVS 2010 emphasised validity.

The various benchmark tests were carefully constructed to explore the performance of EVS 2010 in a number of contexts and under varying conditions. The ideal would have been to assess its performance in real NOP trip setpoint applications, but it would not have been possible to validate EVS 2010 properly using real applications because the true values of the reference TSP would not have been available. In particular it would not have been possible to assess whether it achieved the required coverage by being below the true t_γ with E-probability at least γ.

The benchmarking tests therefore employed simulated scenarios in which the true values of t_γ were known. These were designed to be increasingly complex, with the final exercise based on a scenario that was as close as possible to a real NOP problem.

Although the benchmarking exercises were formally described in terms of two benchmarking rounds, as specified in the contract, they are better understood as three separate exercises based on three distinct simulated scenarios. The detailed specification of each exercise drew upon experience gained in the review and in earlier exercises, and in particular included tests designed to illuminate issues that had arisen in the previous exercises.

> *Benchmark A*. In the first benchmarking exercise a simplified scenario was used. The underlying problem lacked some key features of the NOP problem; in particular the function f(Q, Φ, Θ) did not require any maximisation or minimisation, a characteristic of the NOP trip setpoint problem which is important from a statistical perspective. Nevertheless, its simplicity meant that a relatively large number of Benchmark A tests could be carried out and so allowed a wide variety of detailed conditions to be explored quite fully.

Benchmark B/MCP. The MCP problem is another scenario that is simpler than the full NOP problem. It does feature maximisation but it has no analogue of flux shapes. It is dealt with in the EVS Report as a simpler case to motivate the solution of the full NOP problem. The justification for this scenario is that it includes an extreme (maximisation) feature but is still sufficiently simple to allow a relatively large number of tests to be carried out.

Benchmark B/NOP. The final benchmarking exercise employed a scenario that was as realistic an approximation of the full NOP trip setpoint problem as possible. As such, conducting tests on multiple simulations of the test cases was a non-trivial computational burden, which meant that this exercise comprised relatively few tests.

Full details of the various benchmarking tests and their results can be found in the appropriate appendices to this Final Report. In the discussion below, I have synthesised the many individual findings in these reports to present a more comprehensive analysis of the performance of EVS 2010.

FORMAT OF BENCHMARK TESTS

Each benchmarking exercise was built on a particular scenario, as discussed above. Each scenario had many parameters that could be varied. For instance, in the Benchmark B/NOP scenario it was possible to vary the standard deviations of (and correlations between) computation errors for the physics codes, weights of the various possible flux shapes, the size of the sample of "observed" ripples and numerous other elements. Each scenario had a base case in which all the parameters were fixed at specific values, and a number of test cases in which the values of one or more parameters were varied from the base case. In most tests, the parameter values used in applying the EVS 2010 method (and comparators, where relevant) were the same as the true values, but in some tests EVS 2010 (and comparators) was required to used deliberately mis-specified parameter values, different from the true values. For convenient reference, the tests in each exercise were grouped into test suites where a particular kind of parameter was varied.

All tests were conducted using $\gamma = \beta = 0.95$.

Each test comprised the following general steps:

1. The true value of $t_{0.95}$ was computed using the specified true parameter values for that test. This was done by Monte Carlo simulation of a large number of sample instances of Q, Φ and Θ from their true A-distributions, computing the true ideal TSP value $f(Q, \Phi, \Theta)$ in each case and setting $t_{0.95}$ to the value at which 95% of simulated ideal TSP values lay above and 5% lay below.
2. Using the true parameter values for that test, including the ripples sample size N, a sample of N ripples were drawn at random from the true A-distribution of Q, Then N sets of physics code errors were drawn at random from their true E-distribution and added to the sampled Q values (to produce the "observed" ripples S). Similarly, (except in the case of the Benchmark B/MCP exercise) for each of the M possible flux shapes physics code errors were drawn randomly from their true E-distribution and added to the true flux shapes. Using the simulated "observed" ripples and the simulated computed flux shapes, and using the assumed values of all parameters (which in the case of mis-specification tests would not all equal the corresponding true values), EVS 2010 was used to compute a tolerance limit W.
3. Step 2 was repeated a large number of times, to produce a large set of computed W values. This is then a Monte Carlo simulation of the true E-distribution of W and is used together with the true $t_{0.95}$ from step 1 to calculate various output measures, such as the non-coverage measure which is the proportion of W values lying above the true $t_{0.95}$.

Note that in Benchmark B/MCP the reference TSP is defined to be the value at which 95% of simulated ideal TSP values lay *below* and 5% lay above, because whereas in the original NOP problem the risk in setting the installed TSP is to set it too high the risk in the MCP problem lies in the tolerance limit being too low.

OUTPUT MEASURES

Four output measures were computed (for each test, for EVS 2010 and for comparators if any).

Mean. The Mean measure is simply the mean value of the large number of simulated tolerance limit values W.

SD. The SD measure is the standard deviation of the simulated W values.

Non-coverage. The Non-coverage measure is the proportion of the simulated W values that exceed the true $t_{0.95}$.

Mean Deficit. The Mean Deficit is a more complex measure to define. For every simulated W value that exceeds the true $t_{0.95}$, the deficit is the proportion of simulated ideal TSP values from step 1 that are exceeded by W. The Mean Deficit is the average of the deficits from all the W values that exceed $t_{0.95}$. Notice that the deficit cannot be less than 0.05, and if no simulated W values exceed $t_{0.95}$ then the Mean Deficit defaults to this minimum achievable value of 0.05 (or 5%).

Note that in Benchmark B/MCP the Non-coverage measure is the proportion of simulated W values that are below the true $t_{0.95}$, while the Mean Deficit measure the deficit for each of those W values is the proportion of ideal TSP values that exceed W. This is in accordance with the definition of the reference TSP in Benchmark B/MCP.

ACCURACY OF OUTPUT MEASURES

As explained above, the tests are performed using a large number of simulations leading to a large number of W values. However, the larger the number of simulations performed the greater is the computational burden. In practice, therefore, particularly for the later benchmarking exercises where each simulation can be computationally demanding, the numbers of simulations were limited by practical considerations. It is important to be aware of the fact that the resulting output measures are in effect estimates of the true values of those measures (which would be revealed by using *very* large numbers of simulations).

The values of output measures arising in the benchmark tests should not, therefore, be regarded as precise. This is important when the output measures are used for performance evaluation.

ASPIRATIONS

Before discussing how these output measures are used to formulate performance criteria, it is useful to identify performance aspirations for each measure. Considering Non-coverage first, notice that a tolerance limit method that fulfils the theoretical tolerance limit property should have Non-coverage of $1 - \beta = 0.05$. So the aspiration is for Non-coverage to be equal to 0.05. In practice, because of the various compromises discussed above, we cannot expect Non-coverage to be exactly 0.05. But Non-coverage above 0.05 signifies that the tolerance limit in practice does not provide the level of risk protection that it should have in practice, whereas Non-coverage below 0.05 indicates that it provides more protection than required.

So the practical aspiration for Non-coverage is to be as close as possible to 0.05, but not higher.

We can identify performance aspirations for the other output measures based on what might in principle be achievable given a large quantity of data and physics codes that are very accurate (and so have very small errors). In such circumstances, the large quantity of data and high quality physics codes imply that E-uncertainty will be very small. Then we would expect a good statistical method to be able to yield W values that are close to the true $t_{0.95}$ in every Monte Carlo simulation. Then the Mean would be very close to $t_{0.95}$, the SD would be very small and the Mean Deficit would be very close to its minimum possible value of $1 - \gamma = 0.05$.

The remaining aspirations are therefore that the Mean should be as close as possible to the true $t_{0.95}$, the SD should be as small as possible and the Mean Deficit should be as close as possible to 0.05.

ABSOLUTE PERFORMANCE CRITERIA

The output measures are used to benchmark the performance of EVS 2010. For this purpose a number of performance criteria were defined in each benchmarking exercise. Although these differed in detail from one benchmark exercise to the next, the basic principles are the same in each case.

When applying the performance criteria it is important not to declare that EVS 2010 has failed some criterion on the basis of the computed value of the output measures in the benchmark tests, because these values are not precise and the apparent failure may simply be due to chance. A failure of a criterion can only be asserted when the value of the relevant measure is *significantly* outside the acceptable range for that criterion. Then, even allowing for simulation error, we can be confident that the criterion would be failed by the true value of the output measure.

We first consider criteria that are appropriate for absolute benchmarking, also known as validation.

TOLERANCE LIMIT CRITERIA

The first two criteria concern how well a method performs against the theoretical tolerance limit property.

> **Non-coverage Protection Criterion.** The Non-coverage measure should be less than or equal to 0.05.

As already discussed, in theory the Non-coverage should be exactly $1 - \beta = 0.05$, but we acknowledge that in practice it will not be an exact tolerance limit in this sense. As long as the Non-coverage does not exceed 0.05 then the method achieves at least the level of protection specified in the tolerance limit parameter $\beta = 0.95$.

If EVS 2010 meets the Non-coverage Protection Criterion in every test then clearly its fitness for purpose cannot be criticised on the tolerance limit aspect of performance.

> **Non-coverage and Mean Deficit Criterion.** In any test for which the Non-coverage Protection Criterion is not met, the Non-coverage should not be excessive and also the Mean Deficit should not be excessive.

The criterion uses the word “excessive” twice, but does not define what levels would be excessive. The judgement of fitness for purpose is always qualitative. In the context of the NOP trip setpoint problem, safety is the prime consideration for CNSC, and the decision regarding whether any degradation of tolerance limit performance is excessive must be a judgement for CNSC to make.

Recognising that a method is in theory aiming to achieve coverage of β exactly, in practice it may be reasonable to allow for its non-coverage to exceed 0.05 in some tests. This would indicate that the level of protection that in theory is guaranteed by the tolerance limit is not always achieved, but as long as the Non-coverage is not far above 0.05 this may not be considered enough to render it unfit for purpose.

However, the Mean Deficit measure is another aid for assessing fitness for purpose. On occasions when the tolerance limit W exceeds $t_{0.95}$, the deficit quantifies how serious might be the consequence of setting the installed TSP equal to W. When W equals $t_{0.95}$ this means that the trip will correctly operate early enough on 95% of the events described by the A-uncertainties. When W exceeds $t_{0.95}$ the trip will operate early enough less often. If the deficit is 10%, for instance, then if we set the installed TSP equal to W it will operate early enough on only 90% of those events. The logic behind the tolerance limit is that as long as Non-coverage is less than or equal to 5% we do not worry about the value of the Mean Deficit because the deficit arises only on those 5% of events. However, if Non-coverage is larger the consequences in terms of Mean Deficit become more important.

The Non-coverage and Mean Deficit Criterion also requires that Mean Deficit should not be excessive whenever Non-coverage is more than 5%. The combination of the two elements of the criterion ensure both that performance does not depart excessively from the theoretical 5% Non-coverage and that the Mean Deficit is not so large as to make the potential consequences of those departures excessively serious.

In practical evaluation of fitness for purpose, Non-coverage and Mean Deficit should be considered together.

Finally, note that these criteria apply to all tests in which there is no mis-specification, and also to tests where mis-specification is modest, so that mis-specification to that extent would not be implausible in practical applications. I suggest that the judgement of what level of mis-specification would be plausible in practical NOP trip setpoint applications, for any given parameter, is a matter for agreement between CNSC and the industry.

Meeting the tolerance limit criteria under conditions of moderate mis-specification is a desirable robustness property.

FACE VALIDITY CRITERIA

Statistical reasoning can dictate that a sound statistical procedure for a given problem should have specific properties. Criteria which demand that a proposed method should have these properties are called face validity criteria.

Three face validity criteria apply to a tolerance limit method for the NOP trip setpoint problem. They all concern how the solution W should behave as one of the parameters is varied, so they are assessed by comparing the output measures for two or more tests in the same benchmarking exercise.

> **Sample Size Increase Criterion.** If the size N of the sample of ripples "observations" increases then the Mean output measure should get closer to $t_{0.95}$.

The reasoning here is that increasing quantities of data should allow the statistical method to be more accurate. In the case of a statistical tolerance limit, this means that the limit should on average be able to get closer to the reference value. If this were not the case, then in the real NOP application there would be no incentive to obtain additional data, because on average that would mean having to set the installed TSP lower. Indeed, there would be a perverse incentive to throw data away in order to reduce the sample size!

> **Uncertainty Reduction Criterion.** If the magnitude of standard deviations for any E-uncertainties reduces then the Mean output measure should get closer to $t_{0.95}$.

The reasoning behind this criterion is essentially the same as the previous case. Less noise in the data should mean that the statistical tolerance limit will on average be closer to the reference value. If this were not the case, there would be a perverse incentive to use poorer quality physics codes.

The two cases of increasing sample size and decreasing physics code errors have the clear implication of allowing stronger statistical inferences, but there are many parameters that can be varied and some variations will also have the effect of the Mean output measure moving closer to the reference value. The third face validity criterion applies for all such situations.

> **Mean and SD Consistency Criterion.** If changing a parameter in a particular direction leads to the Mean output measure moving closer to $t_{0.95}$ then the SD output measure should decrease.

There are two ways to argue for this criterion. First, this is consistent with the discussion of aspirations for the output measures. The Mean moving closer to the reference value indicates generally stronger information, which should be accompanied with a reduction in SD. The other explanation derives from the tolerance limit property. If the Mean moves closer to $t_{0.95}$ and the SD also increases then, unless the shape of the E-distribution of W also changes appreciably, these two movements must be accompanied by an increase in Non-coverage. So failure of this criterion indicates poor tolerance limit performance. Even if it does not lead to excessive Non-coverage within the range of parameter variation tested in the benchmarking, there is potential for Non-coverage to become excessive for more extreme parameter values.

Note that for failure of this criterion to be established it is necessary for the SD to increase each time through more than one change of the parameter in question.

Failure of the Mean and SD Consistency Criterion has been described in the benchmarking reports as 'paradoxical behaviour'. Behaviour of this kind had been observed in EVS 2010 applications and noted in my review report. Some of the tests in the various benchmarking exercises were selected specifically to explore the extent of Mean and SD inconsistency.

These criteria apply to all tests in which there is no mis-specification. I believe that the Sample Size Increase Criterion should also apply to tests where mis-specification is modest.

RELATIVE PERFORMANCE CRITERIA

Comparison of the performance of EVS 2010 with that of a comparator can be based on all of the preceding considerations.

First we can consider the aspirations for good performance on each of the output measures. One method can be considered to perform better than another on the Mean measure if its Mean is closer to $t_{0.95}$, on the SD measure if its SD is smaller, and on the Mean Deficit measure if its Mean Deficit is smaller.

The Non-coverage measure allows a variety of comparisons. If one method has a Non-coverage above 0.05 then another method clearly performs better if its Non-coverage is lower. If both are below 0.05 then either may be thought to have performed better. The one with lower Non-coverage has provided more protection in the tolerance limit sense. On the other hand Non-coverage is supposed to equal the theoretical value of 0.05, and from this perspective the method with higher Non-coverage is closer to that aspiration. In general, whereas very low Non-coverage is safe it indicates a method that is making poor use of the available data. So a method which is closer to the theoretical Non-coverage of 0.05 may be deemed a better, more efficient statistical tolerance limit method.

Two methods may also be compared on how well each one satisfies the various other performance criteria – the Non-coverage and Mean Deficit Criterion, the Sample Size Increase Criterion, the Uncertainty Reduction Criterion or the Mean and SD Consistency Criterion.

BENCHMARKING RESULTS

The results of the various benchmark tests are set out in the three reports in the appendices, where the values of the four output measures are tabulated for all the tests and all the methods under test. I will summarise here the principal findings regarding the performance of EVS 2010 against the performance criteria.

NON-COVERAGE PROTECTION CRITERION

This criterion requires that the Non-coverage measure should be no more than 0.05 in all tests except those with mis-specification that is more than would be regarded as modest.

In Benchmark A, this criterion was failed in several tests. Recognising that the Benchmark A scenario was highly artificial and simplified, my report noted two indications that should be explored in the more realistic scenarios of Benchmark B. The first was the occurrence of Non-coverage values above 5% in tests where the flux shape distribution was varied, and the second was the tendency of Non-coverage to increase with sample size. There were no failures of note in tests with modest mis-specification.

In Benchmark B/MCP, several failures of this criterion were noted in tests with no mis-specification, but no pattern was seen to suggest particular situations of concern. The tendency of Non-coverage to increase with sample size was not repeated in the MCP benchmark scenario. The MCP problem does not have a flux shape variable, so this concern could not be explored further in this exercise. In tests with modest mis-specification, some instances were observed of high Non-coverage, reaching almost 100% in one case. These were tests in which the standard deviations of some E-uncertainties were assumed to be higher than their true values.

In Benchmark B/NOP, no failures of the criterion were found, except in some tests where one parameter was inadvertently set at an unrealistic value. In particular, Non-coverage remained well below 5% on all tests in which the flux shape distribution was varied, and also in tests where E-uncertainty standard deviations were slightly mis-specified. Although it was again noticeable that Non-coverage increased slightly with sample size, it was always well below 5%. The only test giving rise to any concern as N increased was test 2.1 where one E-uncertainty standard deviation was assumed to be 25% too large and the Non-coverage rose to 2.3% at N = 500.

The NOP benchmark scenario, being the most realistic for NOP trip setpoint applications, is the most important for all the performance criteria. But the necessarily limited scope of tests in Benchmark B/NOP means that we should not ignore indications from the other testing exercises. In general, it seems that EVS 2010 performs well against this most fundamental of tolerance limit criteria. There appears to be an inbuilt tendency for EVS to give very low Non-coverage; in very many tests it was estimated to be zero, meaning that in thousands of simulated sets of data it did not once produce a computed TSP higher than the reference value. For convenience I will refer to this tendency as a 'bias', without meaning to imply that EVS 2010 is biased in an formal statistical or other sense. It is this 'bias' which appears to be responsible for the method's robustness to modest mis-specification, and in general to its good Non-coverage Protection performance.

The 'bias' must arise from the way EVS 2010 is implemented, rather than from the theory. The reason behind it has not been fully elucidated by AMEC NSS and so there is no assurance that it will persist in all applications. In particular, the finding in the benchmarking tests that assuming E-uncertainty standard deviations to be larger than their true values leads to raised Non-coverage has apparently been seen also in AMEC NSS's own testing. Furthermore, the way that Non-coverage often increases with sample size suggests that the 'bias' has less force for larger N. Whilst true values can never be known in practice, it is clear that every effort should be made to ensure that the magnitudes of physics code errors should not be over-estimated in real NOP trip setpoint applications.

It should be noted that this concern may possibly apply also to the assumption of superposition and independence. AMEC NSS have suggested that the assumption may be conservative because it implies more overall uncertainty, yet it is precisely when assuming more uncertainty than is present in reality that EVS 2010's Non-coverage performance appears to be most unreliable. Nevertheless, it is important to note that the assumption implies increased A-uncertainty, whereas the performance issues referred to have been found only when mis-specifying E-uncertainties. It is therefore not clear whether superposition and independence is a safe or unsafe assumption in practice.

Subject to this important caveat, to the extent that it has been tested in the most realistic Benchmark/NOP scenario EVS 2010 has very satisfactory performance on the Non-coverage Protection Criterion.

NON-COVERAGE AND MEAN DEFICIT CRITERION

This second tolerance limit performance criterion allows some flexibility in practice when the primary Non-coverage Protection Criterion is not met. It requires only that Non-coverage should not be "excessive" and that Mean Deficit should also not be "excessive". So values of Non-coverage above 5% are not regarded as serious failures in performance as long as they are not too high and also do not give rise to large Mean Deficit values. What is "excessive" in both cases for NOP trip setpoint applications should be determined by the competent authorities, but in the benchmark exercises I have applied my own personal judgements in order to provide some tentative conclusions.

In Benchmark A, Mean Deficit was always found to be small. In almost all instances it was less than 6%, which is very close to the theoretical limit of 5%.

The Benchmark B/MCP scenario proved to be more critical in the sense that instances were observed of unacceptably high values of Mean Deficit. In two tests, one with no mis-specification and one with mild mis-specification, both Non-coverage and Mean Deficit were excessive in my judgement. In test 1.4, which has no mis-specification, with a sample size of N = 100 the Non-coverage was 14.2% and the Mean Deficit 29.3%. However, Mean Deficit seems to reduce generally with sample size and no unacceptable results were found in any test with N = 500.

In Benchmark B/NOP the same potential for high Mean Deficit values was not observed. It never reached 10% in any of the benchmark tests.

The different benchmarking exercises provide conflicting information about Mean Deficit. In the MCP scenario high values of Mean Deficit were observed in several tests, while in Benchmark A the Mean Deficit was always small. The Benchmark B/NOP exercise fell between these extremes. In general I do not feel that Mean Deficit has proved to be useful in mitigating the importance of Non-coverage. If Non-coverage is excessive then Mean Deficit cannot be relied upon to be small enough to provide acceptable performance with respect to the Non-coverage and Mean Deficit Criterion.

The conclusion remains that EVS 2010 generally has sufficient inbuilt tendency to low Non-coverage (its 'bias') to give compliance with the tolerance limit performance criteria, but that there are two areas of concern. The first is when standard deviations of E-uncertainties are assumed to be higher than the true values, even by modest amounts; EVS 2010 cannot be said to perform satisfactorily in practice unless the risk of this kind of mis-specification is seen to be very small. The second is when the sample size (i.e. the number N of "observed" ripples) is large, because it seems that the 'bias' may weaken with increasing N.

SAMPLE SIZE INCREASE CRITERION

This criterion requires the Mean output measure to move closer to the reference TSP as the sample size increases. EVS 2010 satisfied this requirement in every test in all the benchmarking exercises, the only exceptions being tests in which there was substantial mis-specification.

UNCERTAINTY REDUCTION CRITERION

This is a closely related criterion which states that if the magnitude of E-uncertainty decreases then the Mean output should move closer to the reference value. This is only required in tests where there is no mis-specification.

In Benchmark A, some failures of EVS 2010 against this criterion were observed. In two instances, decreasing the standard deviation of E-uncertainty in the physics code estimates of flux shapes caused the Mean to move further from the reference TSP. This occurred for all three sample sizes and I reported this as a significant failure of face validity in the simplified Benchmark A scenario, which required further investigation in subsequent benchmarking exercises.

The Benchmark B/MCP scenario did not offer an opportunity to check that finding because it did not include a flux shape component. However, in these tests there was at least one instance (not explainable by chance) of the Mean moving further from the reference value when the standard deviation of the common error component of E-uncertainty in the physics code estimates of ripples decreased. This represented another violation of the Uncertainty Reduction Criterion, although it was not a consistent finding. In at least one case where the common error standard deviation reduced the Mean did move closer to the reference value.

In Benchmark B/NOP, the criterion was again failed for some test comparisons. In this case, it was decreasing the standard deviation of the common error component of E-uncertainty in flux shape estimates that led to the Mean measure moving further from the reference value.

These results will be discussed in conjunction with the Mean and SD Consistency Criterion,

MEAN AND SD CONSISTENCY CRITERION

This criterion demands that when changing a parameter in a particular direction leads to the Mean moving closer to the reference $t_{0.95}$ then the SD should decrease. It applies only when there is no mis-specification. The Mean and SD Consistency Criterion extends the previous two criteria because under circumstances of sample size increase or E-uncertainty reduction we always have a reduction in the SD output. This is observed in all relevant tests in all three benchmarking exercises. So the failure of EVS 2010 to meet the Uncertainty Reduction Criterion is accompanied by a failure also on the Mean and SD Consistency Criterion.

In the first two benchmarking exercises, the parameter variations were limited to changes in sample size and in standard deviations of E-uncertainties which are covered in the Mean Increase and Uncertainty Reduction Criteria. However, in Benchmark B/NOP a new failure for EVS 2010 against the Mean and SD Consistency Criterion was observed. This arose when correlations between E-errors were varied. As the degree of correlation between errors induced by the physics code responsible for flux shape estimation was increased the Mean output moved closer to $t_{0.95}$ but the SD increased. The changes are consistent over two increases in correlation and cannot reasonably be explained by chance.

In the three benchmarking exercises, departures from the Mean and SD Consistency Criterion have been found in three different types of variation of standard deviations (affecting both flux shape and ripples errors) and in changing correlations. They have occurred in Benchmark A where there are no extremal operations in the f(Q, Φ, Θ) function and in the scenarios that do have these operations and are closer to real NOP trip setpoint problems. All of these failures of EVS 2010 performance provide the potential for adversely affecting performance against the primary tolerance limit criteria, because in principle if the SD continues to increase while sufficiently large changes in the relevant parameters are made then the Non-coverage must be expected to become unacceptably large. The key question, which cannot be answered properly in the benchmark tests, is whether this potential for poor tolerance limit performance would arise in practice for realistic parameter values. The generally good performance of EVS on the Non-coverage Protection Criterion, due to its overall tendency towards very low Non-coverage values, suggests that to the extent explored in the benchmarking exercises this potential is not realised. However, it is clear that EVS 2010 does have behaviour in practice that has been referred to as paradoxical because it violates the Mean and SD Consistency Criterion. Therefore, failing clear explanation from AMEC NSS of this behaviour, any practical application of EVS must be accompanied by well-supported reasons to suppose that it will have adequate Non-coverage in that specific application.

Finally, although adherence to this criterion is not strictly required in the case of mis-specification, it should be noted that there are several such instances in the benchmarking exercises where paradoxical movements arise. These all have the potential to lead to unacceptable Non-coverage if parameters are changed further, although this will in many cases not be of practical interest because it would require substantial and unrealistic levels of mis-specification. One example is worthy of mention because it gives additional support to the concern mentioned earlier regarding the assumption of superposition and independence. This is found in the comparison between tests 1.8 and 2.5 in Benchmark B/NOP. In both of these tests independence is assumed between some groups of E-errors, but in test 2.5 the true correlation is non-zero. Increasing the underlying true correlation while assuming independence causes both the Mean and SD to increase, contrary to the Mean and SD Consistency Criterion. Although this is observed only in one change of parameters, and does not in itself lead to a problem with Non-coverage, it does lend some support to concern over the specification of correlation. In general the correlation structure of E-uncertainties may be much more complex than any of the examples studied in the benchmarking exercises (or, as far as I am aware, in any other tests that have been carried out with EVS 2010). The potential for poor tolerance limit performance under changes in correlations, and particularly in mis-specification of correlations, has not been adequately studied. Therefore the fitness for purpose of EVS 2010 in such conditions is not at all clear.

RELATIVE PERFORMANCE

In Benchmark A, a Bayesian comparator was constructed against which the performance of EVS 2010 was assessed. In terms of the aspirations for the various output measures, the Bayesian comparator generally performed better than EVS 2010 in tests without mis-specification. Its Mean was in almost every test closer to $t_{0.95}$ and its SD was almost always smaller, both indicators that the comparator made better use of the available evidence. Its Non-coverage was also almost always closer to the theoretical 5%, and although it also exceeded 5% in more tests there were few instances when it was larger than 10%, all with small values of N. The comparator does not have EVS's 'bias' and so it is more sensitive to mis-specification. Moving to the performance criteria, the Bayesian comparator out-performed EVS 2010 in respect of face validity criteria, meeting all three of those criteria.

The Bayesian comparator in Benchmark A was not intended as a competitor for EVS 2010, and its relative success in those tests stems at least in part from the fact that it was only designed for the Benchmark A scenario. However, it demonstrates that there is some potential to perform better than EVS 2010, at least in simple problems. In particular, it demonstrates that a valid statistical tolerance limit method does not have to have the face validity failures that EVS 2010 has shown.

For the Benchmark B/MCP problem a more complex Bayesian comparator was devised, but it performed appreciably worse than EVS 2010 in the MCP scenario. It is clear that in more complex problems a quick, simplistic solution is unlikely to perform as well as EVS 2010, which has been developed and refined over several years.

Also in Benchmark B/MCP several other comparators were employed. These were variations on the basic "traditional" approach that was in use before EVS 2010 was proposed. The essence of this approach is that where there are "observations" that result from physics codes the underlying true values could be inferred statistically by "subtracting" a random error. For instance, if a physics code produces a value x which can be viewed as a true value z plus a computing error e, then the true value z is the observed x minus the error e. We do not know what e is, of course, but we know (by assumption, at least) its probability distribution. So if e is, for instance, normally distributed with zero mean and standard deviation s then z is equal to x minus a normally distributed random variable with mean zero and standard deviation s. One way to obtain the implied distribution of z is by Monte Carlo, i.e. sampling many e values from the assumed normal distribution and subtracting them from the observed x. The "traditional" approach is based on this way of making statistical inference about the underlying true values being estimated by physics codes.

Three variations that were tested in Benchmark B/MCP shared this basic approach and analysed the Monte Carlo samples in different ways to produce a computed TSP, while the fourth variation simply treated the observed values as true. None was theoretically a proper tolerance limit method. All four methods performed, not unexpectedly, quite differently from EVS. Where EVS has a tendency to produce very low Non-coverage, a 'bias' that is associated with Mean output being relatively far from the reference $t_{0.95}$, all of the "traditional" variants had this 'bias' to an even more marked extent. This rendered them substantially inferior to EVS 2010, even though they all satisfied the Non-coverage Protection Criterion by dint of having Non-coverage zero in every test. None of these methods failed the Uncertainty Reduction Criterion, and so in this respect performed better than EVS 2010, but the primary conclusion from this is not that they are superior methods but that, again, the problems that EVS 2010 has with failing some face validity criteria are not necessarily intrinsic to all statistical approaches.

LIMITATIONS OF BENCHMARKING

Before drawing final conclusions from these results, it is important to interpret them in the light of the limitations of benchmarking.

In the application of EVS 2010 we can identify two different groups of factors that might affect its performance.

1. *Mis-specifications.* Assumptions, models and parameter values that are required to be specified in order to apply the method may be mis-specified.
2. *Inaccurate computation.* Approximations are made in both the theory (the central limit theorem approximation) and the application (particularly the surrogate technique) of the method that result in the computed tolerance limit W being inexact.

The implications of the two groups are assessed in different ways in the benchmarking. Mis-specifications are handled explicitly by tests in which particular parameters are mis-specified. The limitations of this assessment are simply that we cannot explore all the effects of all possible mis-specifications.

- The assumption of superposition and independence was made throughout and no tests were run in which it did not hold. It is possible that this assumption is not true.

- Known distributional forms (typically lognormal) for E-distributions of physics code errors were specified throughout and no tests were run in which these were mis-specified. In my judgement, this kind of mis-specification is unlikely to affect performance substantially but that is simply a judgement, not based on benchmarking results.
- Known distributions were specified for flux detector errors Θ and no tests were run in which these were mis-specified. In practice these distributions are based on empirical data which, even if numerous, would still admit the possibility of mis-specification.
- Other possible mis-specifications were explored by varying parameter values. However, the range of variation was necessarily limited, and rarely were two or more parameters mis-specified in the tests. To the extent that degrees of mis-specification considered in the tests are not representative of, or do not cover the reasonable range of, mis-specification that may be plausible in practical applications, the benchmarking tests may fail to assess fully the impact of these mis-specifications.

Inaccurate computation effects are assessed implicitly in the benchmarking tests where there is no mis-specification. To the extent that performance of EVS 2010 does not match the theoretical tolerance limit requirements in such tests, this is due to approximations made in the theory or application. In this way, all of the approximations are fully explored, subject only to one further limitation.

- The effect of approximations is assessed in all the correctly specified benchmarking tests, but may not be a good guide to the effect of those approximations in practice because the benchmarking test scenarios are not a perfect representation of real NOP trip setpoint problems. Benchmark A and B/MCP tests are obviously not representation of NOP problems, although the insights gained in these tests are important indications. The Benchmark B/NOP scenario is about as realistic as it is possible to be, but it is necessarily imperfect because the “true” ripples and flux shapes are actually observed outputs of physics codes. Also, it only concerns a particular reactor configuration.

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

POSITIVE FINDINGS

1. EVS 2010 is a theoretically sound tolerance limit solution to a range of problems including the NOP trip setpoint problem. I firmly believe that tolerance limits provide the correct statistical framework for risk management situations such as the NOP trip setpoint problem, and since I am not aware of any other proposed approach to this problem that is formulated as a tolerance limit, EVS 2010 appears in this sense to be the only available solution.
2. EVS 2010 as applied, and in particular including the development set out in the EVS 2011 Report, in the great majority of benchmarking tests, and in particular in the group of tests designed to be closest to real NOP trip setpoint applications, maintains a non-coverage no higher than the theoretical 5%. Some important exceptions and caveats are set out in the following section on negative findings, but in general the performance of EVS 2010 on the primary tolerance limit criteria has been satisfactory to the extent revealed in benchmark tests (and subject to the limitations of those tests).

NEGATIVE FINDINGS

3. The EVS 2010 theory is a rather complex development of a tolerance limit. This complexity makes it difficult to understand the mechanisms by which certain kinds of behaviour arise.
4. Like any statistical method, the EVS 2010 theory relies on a number of assumptions which might not hold in practice. In particular the assumption of superposition and independence is acknowledged to be at best an approximation.
5. The EVS 2010 theory assumes that the structures of error distributions are known. There is potentially great complexity in the variance and covariance structures for the error vectors, whereas in practice rather simple structures (equal variances, equal or zero correlations) have been assumed. The benchmarking suggests that the details of these structures may be important, but little is known about the behaviour of EVS 2010 under variations and mis-specifications of these structures.
6. The EVS 2010 theory also assumes that the values of various constants are known. In practice these will not be known and can at best be estimated. A so-called surrogate methodology is employed for some of these, which involves several levels of approximation. The EVS 2011 Report introduces a calculation of the variance of a key part of the trip setpoint computation arising from estimation errors, but this is then used only to make an ad hoc adjustment which does not properly account for the uncertainty introduced by estimation.
7. In practice, EVS 2010 appears to be very sensitive to mis-specification of the standard deviations of computational errors in physics codes. Even a modest over-estimation of these standard deviations can lead to unacceptable levels of non-coverage.
8. The benchmarking exercise has revealed a variety of instances of what has been called paradoxical behaviour in some of the appendices to this report but is here referred to as failures to meet the Mean and SD Consistency Criterion. All such instances raise the potential for excessive non-coverage. The positive finding 2 holds because excessive non-coverage has not generally been found in the benchmarking tests, but the potential remains if parameters are set in practice outside the range explored in the benchmarking exercises. The fact that one such instance arises when varying correlations in E-distributions of physics code errors enhances the importance of negative finding 5.
9. EVS 2010 in practice seems to have a 'bias', specifically a tendency to produce computed TSP values that are unnecessarily far from the reference value and so results in non-coverage values that are in almost all benchmarking tests well below the nominal 5%. This has the beneficial effect of providing a cushion to prevent non-coverage becoming excessive under conditions of mis-specification or

paradoxical behaviour, which is shown in the positive finding 2. However, there are some indications that the 'bias', and therefore also its beneficial effect, may reduce as the size of the sample of ripples "observations" increases. The 'bias' also means that EVS 2010 is processing the available information rather inefficiently.

CONCLUSIONS

Positive findings 1 and 2 are encouraging; they suggest that EVS 2010 is basically sound, at least in theory, and has the potential to provide a practical solution to the NOP trip setpoint problem.

However, the negative findings make it clear that at present there are several obstacles to its use. In particular, negative findings 7, 8 and 9 highlight three specific situations where EVS may not provide adequate protection against the risk of dry-out. Furthermore, the limitations of the generic benchmarking employed in this work mean that the encouraging performance in positive finding 2 may not hold for a particular real application. There are therefore several outstanding concerns regarding the use of EVS 2010 for determining NOP trip setpoints in practice.

Negative finding 8, in particular, means that despite the good tolerance limit behaviour of EVS 2010 in the benchmarking tests we cannot be confident that this will be found in real applications where parameter values differ from those explored in the benchmarking. In any proposed practical application it will be essential to provide assurance that EVS 2010 will deliver good tolerance limit behaviour in that application.

RECOMMENDATIONS

A number of concerns have been raised in this report but it is not formally a part of the "verification and benchmarking" remit to be prescriptive about how those concerns should be addressed. Nevertheless, I believe that the following recommendations for future progress in the NOP trip setpoint problem may be found useful by the CNSC and the industry generally.

TOLERANCE LIMITS AND OTHER METHODS

I have highlighted the need for the industry to understand better the way in which a tolerance limit protects against the risk of dry-out in the NOP trip setpoint problem. In particular, I have pointed to a lack of clarity regarding the treatment of flux shapes.

I have also referred in passing to the lack of clear understanding in the industry about the words *epistemic* and *aleatory*. These terms are useful when they are fully understood and used properly. The distinction between epistemic and aleatory uncertainties is linked to, but not synonymous with, the important separation between E- and A-uncertainties in a tolerance limit.

In my judgement a tolerance limit solution is the most appropriate way for a regulator to ensure that risks are controlled in situations such as the NOP trip setpoint problem. There are other methods, such as what I have referred to as the "traditional method", in use currently or recently which are not of this form. They are superficially statistical but in fact do not correspond to any valid statistical inference procedure. Whether or not the industry determines that an acceptable method should be based on a tolerance limit solution, I believe that regulation should demand methods that are scientifically sound in the way that formal statistical inference methods are. Ad hoc approaches, such as the "traditional method", may behave well in various practical tests but it is even harder to be confident that this good performance will hold in real applications. Furthermore, the nature of the protection that they offer will always be less clear than for a tolerance limit solution.

My first recommendation is that CNSC should embark on a research project to investigate appropriate ways to control risk in the NOP trip setpoint problem. It should seek clarification on tolerance limits and aleatory/epistemic uncertainty with a view to understanding the nature of the risks and possible regulatory controls. The research may involve looking at how such risks are dealt with by other regulators, in the nuclear industry and more widely, but must recognise that much of current regulatory practice is historically ad hoc and may not be founded on sound mathematical and statistical reasoning.

ACHIEVING ASSURANCE OF EVS 2010 PERFORMANCE IN APPLICATIONS

Three ways occur to me for achieving the assurance demanded in my third conclusion.

The first is to engage in theoretical, mathematical investigation of the EVS 2010 algorithm in order to identify exactly why the 'paradoxes' arise and exactly why the 'bias' arises. Having gained that in-depth understanding of the algorithm it may be possible to *prove* that the paradoxical behaviour will never significantly jeopardise the good tolerance limit behaviour. This would not address all of the negative findings, but it would resolve the one that most drives the third conclusion. However, negative finding 3 suggests that this approach is unlikely to be realisable.

The second is to modify the theory and/or the application methodology of EVS 2010. AMEC NSS have in the last few days provided an early draft of such a modification, together with some partial and selective testing to suggest that it may avoid some or all of my negative findings regarding EVS 2010. Such a modification would have to be regarded as a new method, to be evaluated and benchmarked with at least as much rigour as has been devoted to EVS 2010.

A third way, which might allow EVS 2010 to be used for real NOP trip setpoint determination without modification, is for any such application to be accompanied by *empirical support studies.* Tests analogous to those performed in Benchmark B/NOP, but targeted to the context of the particular application, might provide the necessary assurance. They should *inter alia* address the following points.

- If there is concern over the assumption of superposition and independence for this application, robustness of the EVS 2010 method to plausible departures from it should be investigated. [Negative finding 4.]
- The base case should use the available data and reactor specifications for the particular application. [Benchmarking limitation.]
- Evidence should be provided that standard deviations of physics code errors assumed in the application are very unlikely to be higher than the true standard deviations, and some tests run in which they are mis-specified (including minor over-estimation). [Negative finding 7.]
- Realistic assumptions should be made about the correlation structure of physics code errors, including spatial correlations, and mis-specification of these should be explored thoroughly in tests. [Negative finding 8.]
- The size of the sample of ripples "observations" in the tests should match the available data in the application. [Negative finding 9.]
- If there is concern over the assumed A-distribution of Θ, over the weights in the A-distribution of Φ, or over the assumed forms of the E-distributions of physics code errors, robustness to their mis-specification should be tested. [Negative finding 6 and benchmarking limitations.]

My second recommendation is that the companies should give active consideration to the idea of empirical support studies.

EVIDENCE REQUIREMENTS FOR REGULATORY APPROVAL

The suggestion of empirical support studies has to be set within the wider framework of evidence requirements for regulatory approval. It should surely be a requirement for an application to have determined its computed trip setpoint using a method that has been demonstrated to be sound and fit for purpose. However, it is arguable that even when the method has been found to be fit for purpose in a general sense its fitness for a particular application might be more strongly justified by empirical support studies.

I believe that it is therefore worth considering whether a generic suite (generic in the sense of potentially applying whatever method has been used to derive the proposed trip setpoint) of empirical studies might be developed by CNSC, with clear criteria for passing those tests. Not only would this strengthen the regulatory regime but it would also provide more certainty to operating companies as to what is required of an acceptable licensing application.

The development of such a suite could address various issues and compromises that were evident in the benchmarking of EVS 2010, for instance the discreteness of the assumed ‘true’ distribution of ripples and the presence of computation errors in the assumed ‘true’ ripples and flux shapes.

My third recommendation is that CNSC should engage in research to develop a generic empirical test suite to form part of its formal evidence requirements for NOP trip setpoint licensing applications.

APPENDICES

The following Appendices are provided as separate documents.

APPENDIX A. REVIEW OF EVS 2010

“EVS2010Review.pdf”, dated 6 August 2011.

APPENDIX B. REPORT ON BENCHMARK A

“Group A Benchmark Final Report.pdf”, dated 22 February 2012.

APPENDIX C. REPORT ON BENCHMARK B/MCP

“Report_on_Benchmark_B_MCP_problem_Nov_19.pdf”, dated 19 November 2012.

APPENDIX D. REPORT ON BENCHMARK B/NOP

“Report_on_Benchmark_B_NOP_problem_v2.pdf”, dated 10 February 2013.

APPENDIX E. DISPOSITION OF COMMENTS ON THE DRAFT FINAL REPORT

“AOH disposition.pdf”, dated 13 March 2013.

Review of the EVS 2010 Methodology

Professor A O'Hagan

August 6, 2011

Contents

Executive Summary

This report presents my evaluation of the mathematical and statistical basis of the methodology set out in the document "A Genuine '95/95' Criterion for Computing NOP Trip Set-points Using EVS Methodology" by Paul Sermer and Fred Hoppe. That document is report number G0263/RP/008 from AMEC NSS Ltd., dated September 30, 2010. Both the document and the methodology that it proposes will be referred to herein as EVS 2010.

This report is prepared for the Canadian Nuclear Safety Commission (CNSC) under Contract 87055-10-1226 – R396.2 - "Independent Verification and Benchmarking of Statistical Method and Mathematical Framework in OPG/BP 2010 EVS Methodology for Calculation of NOP Trip Setpoint".

The principal findings in this report are presented in Sections 16 and 17 and are summarised here.

The first question that I posed myself in this review was: **Is the theory presented in the EVS 2010 document mathematically and statistically correct?** My first conclusion is that the theory is indeed mathematically and statistically correct.

I further conclude that the EVS 2010 is an ingenious solution to a highly complex and challenging problem, and its approach of using a tolerance limit is appropriate to the context of computing NOP trip setpoints. However, there are some subtle issues regarding the meaning of the two quantities, denoted in EVS 2010 and here by γ and β, which define the precise kind and degree of protection provided by the computed trip setpoint. These should be clearly understood in order for γ and β to be given appropriate values for regulatory purposes, and this is my first formal recommendation.

Recommendation 1. The industry, and in particular the regulator (CNSC), should consider carefully the meanings of the two quantities γ and β in a tolerance interval, with reference to the NOP trip setpoint problem. The interpretation of γ requires careful specification of the random circumstances of the future instances in which the NOP trip is to operate, and in particular the probabilities or weights assigned to different flux shapes. In regard to β, the distinction between the 'confidence' interpretation of a frequentist tolerance interval (used in EVS 2010) and the probability interpretation of its Bayesian analogue is also important.

Although the theory may be correct, the second question that I posed for this review is equally important, if not more so: **Is the EVS 2010 methodology fit for purpose?** My conclusion is that there are a number of issues that should be resolved concerning how well the statistical problem defined and solved in EVS 2010 accords with the real-world NOP trip setpoint problem. In any serious application of statistics there are always compromises because the real-world problem is more complex than any statistical theory. The fitness for purpose question concerns whether any simplifications, assumptions and approximations (generically referred to in this review as compromises) made by EVS 2010 render the computed trip setpoint not sufficiently accurate for the purpose of neutron overpower protection.

The next two formal recommendations concern issues that I have identified regarding the accuracy of EVS 2010 computations in the real-world context of the NOP trip setpoint problem.

Recommendation 2. The impact of additional uncertainties that arise implicitly or explicitly from compromises in the EVS 2010 methodology should be assessed through benchmarking tests. Particular attention should be paid to the uncertainty in the flux shape distribution and in the estimate of the factor denoted in EVS 2010 by $r_{1-\gamma}$. Judgement regarding the fitness for purpose of EVS 2010 should be reserved until these tests have been conducted and evaluated.

Recommendation 3. The paradoxical reported behaviour of the EVS 2010 trip setpoint when uncertainties, or estimates of uncertainties, are reduced should be investigated and the paradox resolved, to determine whether it is the methodology (or its implementation) or common intuition which is at fault, or whether it is simply the result of a misunderstanding. Benchmarking tests have a useful role to play. This exercise is also important to carry out before making any fitness for purpose judgement on EVS 2010.

The third question that I posed for this review is: **What other approaches exist and what are their relative merits?** I do not regard earlier methods for NOP trip setpoint computation which do not employ a tolerance limit approach as competitors to the EVS 2010 methodology. However, one of my findings is that the EVS 2010 solution is complex and may be unnecessarily so. In this review I propose an alternative approach based on a Bayesian analogue of the tolerance limit. I believe that this alternative method could be feasible and potentially a simpler and more direct solution. At least it will provide a comparator for EVS 2010 in benchmarking tests. This is therefore my fourth formal recommendation.

Recommendation 4. A more direct, Bayesian approach outlined in Section 14 of this review should be contrasted with EVS 2010 in benchmarking tests.

My fourth and final component of the scope of this review was the question: **Are there other mathematical or statistical methodologies that could contribute to improve the EVS 2010 methodology?** I have made some small suggestions, but overall I do not press for any of these to be considered prior to the resolution of the issues raised in the above four recommendations.

Finally, although I make only four formal recommendations, it would be a mistake to focus only on these. I believe there are many more points in this review which, although not so important in my judgement that I make them formal recommendations, would nevertheless repay careful reading.

1 Context, scope and structure of this review

This report is prepared for the Canadian Nuclear Safety Commission under Contract 87055-10-1226 – R396.2 - “Independent Verification and Benchmarking of Statistical Method and Mathematical Framework in OPG/BP 2010 EVS Methodology for Calculation of NOP Trip Setpoint”. A key task under that contract is to review the document “A Genuine ‘95/95’ Criterion for Computing NOP Trip Set-points Using EVS Methodology” by Paul Sermer and Fred Hoppe, which will be referred to herein as EVS 2010. The contract’s formal specification of the task is:

> Perform a review of the EVS 2010 mathematical framework as proposed by OPG and BP for the calculation of NOP trip setpoints with a focus on technical soundness of its mathematical and statistical basis in relation to NOP problem formulation and physical simulation methodology.

During the startup meeting in Ottawa on April 6th, 2011, I expanded this to offer four questions for the review to answer.

1. Is the theory presented in the EVS 2010 document mathematically and statistically correct?
2. Is the EVS 2010 methodology fit for purpose?
3. What other approaches exist and what are their relative merits?
4. Are there other mathematical or statistical methodologies that could contribute to improve the EVS 2010 methodology?

The first two questions are fundamental, and the conclusion of this review must inevitably be negative if either question does not have a clear, affirmative answer. The first simply asks whether there are any errors in the mathematical or statistical analysis. If the answer to this question is ‘Yes’, then we can conclude that the EVS 2010 method, if correctly implemented, will provide a correct solution to whatever problem the method addresses. The EVS 2010 theory involves much complex statistical analysis, as well as some conceptual complexity in terms of representation of uncertainty and the idea of tolerance limits.

The second question asks to what extent the problem which EVS 2010 addresses is an accurate reflection of the NOP trip setpoint problem. Any statistical analysis is based on a formal mathematical expression of the real-world problem to be solved, and this is almost inevitably a simplification of that real-world problem. Assumptions and approximations may be made, and we must ask whether the solution to the formal mathematical problem will be acceptably close to the desired solution of the real-world problem. It is important to identify any assumptions, approximations and simplifications made in the EVS 2010 methodology in order to decide whether the methodology is fit for the purpose of computing NOP trip setpoints.

The third and fourth questions ask whether we can do better, either with an alternative approach altogether or by improving on the presented approach.

My review has concentrated on the EVS 2010 document itself, but I found a number of aspects of the report needed further clarification. So my reading of EVS 2010 has been augmented by various discussions, some supplementary reading and responses to a series of questions that I posed to Ontario Power Group (OPG) and Bruce Power (BP), and thereby indirectly to AMEC NSS. These questions and responses, together with other formal correspondence between myself and OPG/BP, are incorporated in this report as Appendix C .

The review is structured in three parts. Part I provides important starting points for the EVS 2010 approach. It comprises Sections 2 to 6 and addresses the basic NOP trip setpoint equation, the general concepts of tolerance limits and the nature of the flux shape distribution. Sections 7 to 11 comprise Part II and provide a summary of the EVS 2010 approach to setting the NOP trip setpoint. The summary is considerably less detailed than the EVS 2010 document but is intended to present the structure of the approach in a clear progression, and to identify specific aspects that will be significant in this review. Part III presents the findings of this evaluation of the EVS 2010 approach, developed through Sections 12 to 15 and ending with the review conclusions and recommendations in Sections 16 and 17. Key findings are presented in the Executive Summary at the front of this report.

To aid the reader in navigating the document and in following the development, each of the three parts begins with a very brief description of its constituent sections.

Throughout this document a number of important points are highlighted as Remarks; for convenience these are collected together in Appendix A.

Part I
Background

EVS 2010 presents a mathematically sophisticated and complex solution to the problem of computing NOP trip setpoints. I will attempt to review it without going into the mathematical detail but still doing justice to the many important and often ingenious steps involved in the methodology. This first Part of the review comprises five sections which examine some of the most basic concepts. Very briefly, these chapters address the following points.

1. *The ideal NOP trip setpoint.* A fundamental equation in the EVS 2010 document expresses the theoretical ideal value for the trip setpoint if it is to trip in time but not too early.

2. *The actuality distribution.* In any future actual instance when the NOP trip might be activated, the detailed reactor conditions on which the ideal setpoint depends will not be known. The uncertainty about the circumstances of a future actual instance is an important component of the approach. Given a full specification of the probability distribution of those circumstances, it would be possible to compute a quantile trip setpoint that would trip early enough in a sufficiently high proportion of future actual instances.

3. *The tolerance interval framework.* However, the details of that probability distribution, called the A-distribution, will not in fact be known. The concept of a tolerance interval is to use observational data to provide statistical evidence about the unknown features of the A-distribution, and so to derive a statistical confidence limit for the quantile trip setpoint. This is called a tolerance limit. The sampling distribution of the data is called here the E-distribution.

4. *The possible flux shapes.* In the NOP trip setpoint problem, several uncertain components potentially have very complex A-distributions or E-distributions. Of particular importance is the distribution of flux shape and how this is handled in EVS 2010. The first of two sections concerning flux shape examines the set of possible shapes.

5. *The flux shape distribution.* The probilities or weights that are assigned to the possible flux shapes play a major role in determining the trip setpoint, but there is little evidence concerning the less common flux shapes. The distribution is closely linked with the question of how to define the population of actual future instances.

2 The ideal NOP Trip Setpoint

The acronym NOP stands for Neutron Overpower Protection, and concerns a general protection against the power in a CANDU nuclear reactor reaching too high a level. The NOP trip is triggered when flux detector readings in the reactor's safety channels exceed a value known as the NOP trip setpoint (TSP). Specifically, a reactor has a number of safety channels with several detectors in each safety channel, and the NOP trip may be activated according to various criteria regarding which of these detectors have readings exceeding the TSP, but the criterion typically considered in EVS 2010 is that at least one of three detectors must exceed the TSP in each of three safety channels.

The TSP is intended to avoid the power in any of the reactor's fuel channels reaching a level called the Critical Channel Power (CCP). The CCP is in general different for each fuel channel in the reactor, and is a value at which the fuel in that channel becomes in danger of a condition known as dry-out. The purpose of the NOP TSP is essentially to avoid dry-out occurring in any fuel channel.

In order to identify a suitable TSP value it is necessary to establish a logical link from the flux detector readings in the safety channels to the powers in the fuel channels. Ideally, the TSP would be set so that the NOP trip is triggered precisely when one channel power reaches its CCP, but it is not possible to achieve this, for a variety of reasons. One simple reason is that the flux detectors are inevitably subject to measurement errors, although I understand that these are small. A more significant reason is that the reactor's state is continually changing, and in different states the ideal value of the TSP will be different. The state is conventionally thought as being comprised of two components.

1. *Ripples.* CANDU reactors are re-fuelled without the need for shut-down, so there are frequent changes in the fuel in the various channels. Fresh fuel produces higher power in the channel itself with complex interactions on neighbouring channels. Given any history of fuelling actions, the powers in the various channels will deviate from their nominal levels and these deviations are called ripples. Ripples do not change the overall total power of the reactor, but they change the powers in individual channels, making them closer or further from their CCP values. Because the flux detectors are calibrated to total output over a region of high power channels, the TSP depends on the minimum ratio of the CCP to actual channel power in any channel, and this clearly is influenced by ripples.

2. *Flux shape.* The powers in individual channels can deviate also for reasons other than the fuelling history. In particular, this can happen because of incidents or errors in reactor management which cause a slow loss of regulation (LOR). It is these LOR incidents that the NOP trip is intended to protect against, since in an LOR event the total reactor power will rise causing eventual dry-out if there is no recognition of the problem and appropriate response. The flux shape refers to a characteristic signature of the particular kind of incident or error, in terms of factors called channel over-powers (COPs) multiplying the various channel powers (which acts

on top of ripples). It means again that individual channel powers may be closer or further from their CCP values, thereby also affecting the TSP. An additional factor is that flux shape also influences the CCPs.

EVS 2010 defines an *ideal* TSP for any particular reactor state as follows.

$$TSP = \min_i \left\{ \frac{CCP_i}{CP_i \times COP_i} \right\} \times RP^{ind} \times CF \times \min_{CH} \max_k \{FX_{CH,k} \times FOP_{CH,k}\} \ , \tag{1}$$

where

$$CP_i = CP_i^{ref} \times Q_i \times RP \tag{2}$$

and where

- CCP_i is the CCP in fuel channel i,
- COP_i is the COP in fuel channel i, due to flux shape,
- RP^{ind} is the indicated total reactor power that has been used to calibrate the flux detectors, as a fraction of the reactor full power value FP,
- CF is the calibration factor that has been applied to the flux detectors,
- $FX_{CH,k}$ is the nominal flux detector reading at the k-th detector in safety channel CH, excluding the effect of flux shape,
- $FOP_{CH,k}$ is the Flux Over Power at the k-th detector in safety channel CH due to flux shape,
- CP_i^{ref} is the reference channel power in fuel channel i, which would in principle apply under full reactor power
- Q_i is the ripple effect in fuel channel i,
- RP is the actual total reactor power, also expressed as a fraction of FP.

Equation (1) is given in EVS 2010, where it is numbered (5). I will refer to equation numbers and section numbers in EVS 2010 with a prefix 'EVS-', so that equation (1) above is the EVS 2010 equation (EVS-5). Similarly, (2) is (EVS-6). To understand the logic of (1), first note that $FX_{CH,k} \times FOP_{CH,k}$ is the true flux detector reading at the (CH, k) detector. The operation of $\min_{CH} \max_k$ applied to this produces the effective detector reading that is to be compared with the TSP, because of the criterion that at least one detector in each channel must exceed TSP in order to trigger a trip. This is then multiplied by $RP^{ind} \times CF$ which converts the effective detector reading to a figure which is interpreted as a fraction of FP.

The term $\min_i \left\{ \frac{CCP_i}{CP_i \times COP_i} \right\}$ is the ratio of CCP to actual channel power for the channel which is nearest to its CCP (in percentage terms). This is called the *margin to dry-out*. We can now interpret equation (1). Under whatever reactor conditions apply, if the effective detector reading, expressed as a fraction of FP, were to increase by a factor equal to the margin to dry-out, then the channel powers would increase by the same factor and one of them would hit its CCP. This is exactly the condition that we defined as being the ideal point at which the NOP trip should trigger, therefore this determines the ideal TSP.

Remark 1 *The NOP trip setpoint problem as formulated in EVS 2010 and as adopted here focuses on whether the channel power in any fuel channel exceeds its CCP. The ideal trip setpoint will ensure that the NOP trip operates in a slow LOR incident as soon as any channel power reaches its CCP. The validity of EVS 2010 depends in the very first instance on this being an accurate formulation of the NOP trip setpoint problem.*

There is an important assumption implicit in the basic equation (1) if it is to fulfil the fundamental criterion in Remark 1.

Remark 2 *The definition of TSP in (1) or (EVS-5) implies an assumption that in an LOR event all channel powers increase in fixed proportion to the total power, and all flux detector readings also increase in fixed proportion to total power. This assumption has been acknowledged and defended by OPG and BP representatives as at least an accepted approximation to reality. The strength of the assumption is that it allows (1) to be used under any reactor conditions, even when overall reactor power is such that the reactor is nowhere near a dry-out condition, to deduce what TSP should theoretically apply under those conditions if reactor power were to increase due to slow loss of regulation. The EVS 2010 methodology relies on this assumption and to the extent that it is only an approximation this limits the validity of EVS 2010.*

The assumption is implicit in what is known as the *superposition principle*, which represents actual channel power in channel i at any time as the product of four components — the total reactor power (RP), its reference channel power (CP_i^{ref}), its ripple (Q_i) and its channel overpower (COP_i). The assumption in the superposition principle is that these are independent factors, and in a slow loss of regulation event the total power rises while the ripples and flux shape (channel overpowers) stay fixed. There is another implicit assumption that the flux shape characterising the LOR initiating event operates independently of the disposition of ripples at that time. The reality is almost certainly different — the effect of the LOR event on channel powers is not independent of ripples, and nor does it stay constant as reactor power increases. As stated in Remark 2, the superposition principle (and these implications of it) appears to be an acknowledged and accepted position in the context of the NOP trip setpoint problem.

The following quantities in (1) and (2) are fixed and known: $CP_i^{ref}$and the reactor full power FP (which is simply the sum of the CP_i^{ref} values). The values of RP^{ind} and CF can also be considered fixed because they are used in the calibration of detectors and are only in the equation to express the detector readings on the right scale. The other quantities are variable, and as a result the ideal TSP given by (1) is variable.

3 The actuality distribution of TSP

In any future actual reactor conditions, we will not know what ideal TSP value applies because the values of the variable quantities in (1) will be unknown. We therefore regard them as random variables and characterise their uncertainties as follows.

- Q_i is uncertain because the ripples are unknown. We will know the history of fuelling events, but to derive the ripples values from this involves complex physics computations. This is one of the functions of the SORO computer code, but SORO is not used on-line during an LOR event (and like any complex physics code its computations are subject to *model discrepancy* errors).
- COP_i and $FOP_{CH,k}$ are uncertain because the flux shape applying at the time will also be unknown, and because deducing channel/flux overpowers from a given flux shape is another complex physics computation (in this case performed by TUF, based on flux shape computations by RFSP).
- CCP_i is uncertain because it depends on flux shape and on other thermal hydraulic conditions in the reactor.
- RP is uncertain because we can only measure reactor power through a subset of instrumented channels. These are used to calculate RP^{ind}, which is viewed as an estimate of RP.
- $FX_{CH,k}$ is uncertain because detectors drift out of calibration between calibration exercises.

EVS 2010 describes these uncertainties as *aleatory*. In discussion it has been accepted that the use of the words ‘aleatory’ and ‘epistemic’ to characterise uncertainties in EVS 2010 does not necessarily adhere strictly to the way these terms are used in other contexts, but that it is necessary to distinguish between two groups of uncertainties in the methodology. The first group are the uncertainties listed above. We will call them A-uncertainties, where the ‘A’ can be read as ‘aleatory’ or as ‘actuality’ (referring to the actual reactor conditions in a future incident).

Because of these A-uncertainties, TSP is a random variable. Using $\Pr_A$ to denote probabilities induced by the A-uncertainties, if we choose to set the actual NOP TSP at a value t then there is a probability $\Pr_A(TSP < t)$ that in an actual incident this will be too high and the trip will be activated too late, with a consequent risk of dry-out. If we knew the A-distributions of all these uncertain quantities, we could compute $\Pr_A(TSP < t)$ for any t. Then the choice of t becomes a matter of deciding what risk of dry-out is acceptable. For instance, we could choose a probability γ and set $t = t_{1-\gamma}$ where $t_{1-\gamma}$ solves the quantile equation

$$\Pr_A(TSP < t_{1-\gamma}) = 1 - \gamma \ . \tag{3}$$

Setting γ at some high probability (such as 0.98) means that there is only a small (e.g. 2%) chance of not tripping early enough.

Unfortunately, the A-distributions are not known. Let δ denote in general whatever features, aspects or parameters of the A-distributions are unknown. Then the TSP chosen by the above rule becomes a function of δ and we denote it by $t_{1-\gamma}(\delta)$.

4 The tolerance limit framework

The idea of a tolerance limit is to use other evidence to estimate the quantile TSP $t_{1-\gamma}(\delta)$. Formally, we compute a limit T^* from the available evidence. T^* is a random variable in standard frequentist statistical theory and we let $\Pr_E$ denote probabilities relating to the sampling distribution of T^*. The tolerance limit property is given by choosing another probability β and then constructing T^* so that

$$\Pr_E(T^* < t_{1-\gamma}(\delta)) = \beta \tag{4}$$

for all possible values of δ. Then T^* (or more strictly, the method by which T^* is calculated from the data) is a γ/β tolerance limit for the TSP. It has the property that there is 100β% E-probability that when T^* is calculated from the evidence in this way the A-probability of the NOP trip activating before dry-out is at least 100γ%.

This is the idea followed in EVS 2010. The evidence used in the NOP trip setpoint problem is derived from data recorded on historic snapshots of reactor operation. The group of uncertainties around this evidence are called *epistemic* in EVS 2010. However, in a formal frequentist tolerance limit calculation the E-probabilities $\Pr_E(\cdot)$ are actually also aleatory. Regardless of whether we read the 'E' as 'epistemic' or 'evidential' we use the term 'E-uncertainty' to describe the uncertainties in the evidence base. Further clarification of the A- and E-uncertainties, and of aleatory and epistemic uncertainties is given in Appendix B.

In my opinion it is right to separate the A-uncertainties and the E-uncertainties in the NOP trip setpoint problem, and to derive a trip setpoint as a tolerance limit. Some earlier approaches to the problem have not made this separation.

Remark 3 *The tolerance limit approach adopted by EVS 2010 distinguishes between uncertainties relating to future actual reactor conditions (which I term A-uncertainties) from the uncertainties (E-uncertainties) relating to evidence used to learn about unknwown features of the A-uncertainty distributions. This is an important contribution to the NOP trip setpoint problem because it focuses on controlling the proportion of times (i.e. A-probability) that the NOP trip activates sufficiently early in all future actual instances.*

It is worth noting that a Bayesian formulation of the problem would be different and might be preferable. In this case the evidence would be converted to a posterior distribution for δ, and I will use $\Pr_\delta$ to denote probabilities with respect to this posterior distribution. We would then set T^* to solve

$$\Pr_\delta(T^* < t_{1-\gamma}(\delta)) = \beta \ .$$

Although superficially similar to the frequentist tolerance limit, the interpretation is different. The frequentist interpretation is that if we obtained many random sets of data of the kind we actually had (e.g. from many more sets of historic snapshot data) and calculated a T^* from each such set, then $100\beta\%$ of these T^* values would be below the theoretical setpoint $t_{1-\gamma}(\delta)$. However, we cannot say whether the actual T^* we have calculated has this property. We cannot say that there is a $100\beta\%$ probability that *this* T^* is low enough, and indeed in the frequentist framework there is no such probability relating to *this* T^*. We can only say that in the long run of many repetitions $100\beta\%$ of T^* values computed in this way will be low enough.

In contrast, the Bayesian analysis does say that there is a $100\beta\%$ probability that *this* T^* is low enough. And the δ probabilities $\Pr_\delta(\cdot)$ in the Bayesian analysis are genuinely epistemic.

Remark 4 *The EVS 2010 objective is to compute a trip setpoint value T^* as a frequentist γ/β tolerance limit. It is important to recognise that the correct interpretation of this tolerance interval is that if we were able to repeat the computation of T^* many times using new sets of historic data then $100\beta\%$ of these T^* values would be below the value $t_{1-\gamma}$ at which the trip would successfully activate sufficiently early in $100\gamma\%$ of future actual instances. This does not imply that there is a $100\beta\%$ probability that the actual calculated T^* will be below $t_{1-\gamma}$. Such an interpretation can only be given for the analogous Bayesian limit.*

There are some important difficulties in deriving a tolerance limit for the NOP trip setpoint problem, which have heavily influenced the approach followed by EVS 2010. The first of these is the complexity of the A-uncertainties. In particular, the ripples Q_i constitute a high-dimensional random quantity because there is a ripple value for each fuel channel. I presume that these could not be assumed to be even approximately independent — the term 'ripples' suggests a random field in which there are strong local correlations. The ripples arise from complex interactions between fuelling events over time spans comparable to the life of a fuel element. So I imagine that it is hard to propose any credible joint probability distribution for the ripples.

The flux shape is also a high-dimensional random variable, but there is not the same difficulty with regard to specifying a suitable A-probability distribution because a previous report to EVS 2010 has proposed a probability distribution that EVS 2010 uses. The probability distribution for flux shape is considered in Section 6 of this review.

The second difficulty is that in the tolerance interval framework the sampling distribution (the E-distribution) of T^* is assumed to depend only on δ. If it also depends on additional unknown parameters these should be treated as nuisance parameters and the tolerance limit equation (4) needs to hold for all values of the nuisance parameters as well as δ.

5 The possible flux shapes

The flux shape is denoted by φ in the EVS 2010 report. In normal operation the flux shape should be flat, meaning that $COP_i = 1$ for all fuel channels and $FOP_{CH,k} = 1$ for all flux detectors. However, there are various situations that can give rise to a flux shape that is not flat, and these in particular include those situations that can give rise to an LOR event. In broad terms, an extreme flux shape is likely to lead to a greatly reduced margin to dry-out and requires a low trip setpoint to avoid dry-out. Consequently, flux shape is an important factor in this problem.

Flux shape is stated in EVS 2010 to be a discrete random quantity. Each of a discrete set of possible scenarios gives rise to its own characteristic flux shape, which is computed using the RFSP physics code. There may be hundreds of distinct flux shapes, but the number is said to be finite and hence φ is treated as a discrete random variable in EVS 2010. My questioning of the representatives of OPG/BP has shown that strictly speaking φ is not discrete. For instance, a control absorber rod which is not fully inserted could lead to a flux shape which depends on just how far it is inserted. Since the position of the rod is a continuous variable, this scenario leads to a continuous set of flux shapes. In reality, φ is not strictly a discrete quantity.

The OPG/BP response to my questions is that it is enough to consider the extreme cases of complete insertion and complete withdrawal, and it does seem plausible that the flux shape for any degree of partial insertion would lead to a trip setpoint between the values obtained from the two extremes. Putting a discrete probability on the most extreme of the flux shapes in some continuous set should lead to a lower trip setpoint.

Remark 5 *EVS 2010 asserts that the set of possible flux shapes is discrete. The reality is that it is not strictly discrete, but treating it as such is not likely to lead to a too-high value for the NOP trip setpoint.*

Another possibility concerns whether the number of possible scenarios is finite. In my questions to OPG/BP, I asked whether it would be possible to dream up more and more possible scenarios but did not receive an answer. It seems to me that this would in principle be the case, that one could always imagine more failure modes. It may be that the finite set of flux shapes that are used in the application of EVS 2010 is sufficient in the sense that all the most extreme possibilities are included. However, I wonder whether there might be some that have been excluded because they are extremely unlikely. For instance, I can imagine that a combination of several events that individually have flux shapes in the defined set might lead to an even more extreme flux shape in combination (and hence an extremely low TSP) but this is not considered because in practice it is highly improbable. To ignore a case such as this may be sensible but could lead to a too-high trip setpoint value.

Remark 6 *I have not received a convincing argument that the finite set of considered flux shapes is complete in the sense that there are not other scenarios* *which might be imagined. If others are possible and might lead to extremely low ideal trip setpoint values then, no matter how unlikely they are, to exclude them might result in over-prediction of the NOP trip setpoint.*

6 The distribution of flux shape

If we accept that there are a finite, discrete set of possible flux shapes, there remains the question of how to assign an appropriate probability distribution for them. The A-distribution of the ideal TSP, and in particular the value of the quantile $t_{1-\gamma}$, is likely to be sensitive to the distribution of flux shape. If we assign a distribution that gives higher probability to the more extreme flux shapes this will presumably lead to a lower $t_{1-\gamma}$.

Most of the time the reactor is in a well controlled state. Flux shapes that are not flat or nearly flat are therefore rare. If we assign a flux shape distribution that reflects the moment by moment frequencies of flux shapes, then nearly all the probability will go to near-flat shapes, resulting in a high $t_{1-\gamma}$ and a high computed trip setpoint. Then unless γ is very close to 1 we simply will not be protecting against any of the less usual flux shapes that it seems to me should drive the NOP trip setpoint problem. I imagine that the choice of $\gamma = 0.95$ is not intended to permit a 5% risk of dry-out every moment or every day. There would be far too many failures. Instead I imagine the intention is to protect against dryout on 95% of the times when the reactor is not in its usual well-controlled state and the flux shape is much less likely to be flat. Some of my speculation here may be naive but my point is that it is important to define carefully the population of future events that is of concern. This A-population defines the A-distributions, and while the A-distributions of other uncertain quantities such as Q may not be at all sensitive to the choice of population, the distribution of φ will be. Furthermore, the appropriate choice of γ will also depend on the A-population.

There is essentially no data on the occurrence of flux shapes that are not almost flat, with which to estimate relative frequencies. In response to one of my questions, it was said that equal probabilities are given within the categories of abnormal shapes. In the absence of relevant data this is a reasonable default choice but it is important to recognise that there is considerable uncertainty about the true frequencies of such flux shapes.

Remark 7 *The TSP is likely to be sensitive to the choice of probability distribution across flux shapes. In this context it is important to think carefully about the population of future events that determines the A-uncertainties, and to assign γ appropriately to that population. It is also important to recognise uncertainty about the flux shape distribution.*

An alternative way to formulate this is to say that the distribution of flux shape determines the nature of the random future events that we wish to protect against. For instance we may choose to give equal probabilities to the flux shapes in a given category, not because we know or believe that they occur with equal frequency in the population of interest but because we choose to define that population as one in which the flux shapes in this category are equally likely. The flux shape distribution is considered further in Section 13.

Part II
Elements of the EVS 2010 Approach

We will now describe the EVS 2010 approach in sufficient detail for the purposes of setting out the findings of this review. The key elements of the approach are dealt with in the next five sections of this review.

1. *The A-equation.* The mathematics of the approach hinge on two fundamental equations. The first of these will be referred to herein as the A-equation, because it concerns the uncertainties in a future actual instance when the NOP trip might be activated. The A-distribution of the trip setpoint is determined through this equation, and in turn formally determines the quantile $t_{1-\gamma}$.

2. *Reducing δ.* As described in Section 4, in general $t_{1-\gamma}$ depends on the uncertain features of the A-distribution that we denote by δ. This section considers the nature of δ and the way that EVS 2010 reduces the formulation to make δ manageable.

3. *The E-equation.* The second fundamental equation formulates the uncertainties in the data that are available to learn about δ and so complete the second part of the tolerance limit framework of Section 4.

4. *The data.* EVS 2010 uses certain averaging and adjustment of the data described by the E-equation. The result is a set of quantities called $V_1, V_2, \ldots, V_n$ whose mean $\bar{V}$ and standard deviation S_V are the fundamental statistics used to construct the tolerance limit.

5. *The EVS 2010 tolerance limit.* The actual tolerance limit is an expression of the form $\bar{V} + \lambda S_V$, where λ is chosen to satisfy the tolerance limit criterion. EVS 2010 employs some quite advanced theory to derive this.

7 The A-equation

The EVS 2010 approach is characterised by a pair of key equations, the first of which is

$$T = T^0(Q, \varphi) + \vartheta(Q, \varphi, \theta) \ . \tag{5}$$

I will refer to this as the *A-equation* because it relates to the uncertainty around a future actual instance. This equation is (EVS-57) in EVS 2010.

Here, T is $\log TSP$, the logarithm of the ideal TSP value TSP in a future actual instance. Q and φ are the ripples and flux shape applying in that future instance.

$T^0(Q, \varphi)$ is the logarithm of a simplified version of TSP, denoted by $TSP^0(Q, \varphi)$. This is the same as TSP except that

(a) we assume that Q and φ are known, and so COP_i and FOP_i are no longer uncertain;

(b) we replace $FX_{CH,k}$ by its nominal value $dr_{CH,k}$, which will usually be 1 because of calibration;

(c) we replace RP by its estimate RP^{ind}; and

(d) we replace CCP_i by its value computed with the assumed φ and with reference thermal hydraulic conditions.

Then for given Q and φ the error term $\vartheta(Q, \varphi, \theta)$ represents deviation of $TSP^0(Q, \varphi)$ from TSP because of the random deviations in $FX_{CH,k}$, RP^{ind} and CCP_i (due to deviation of thermal hydraulic conditions from the reference values but for given φ). The random variable θ represents these random deviations but ϑ has arguments Q and φ because the actual error also depends on these variables.

The A-uncertainties are expressed in the three random quantities Q, φ and θ. EVS 2010 assumes that they are mutually independent, which seems entirely reasonable if the superposition principle is adopted. If values of these are given, we can compute $T^0(Q, \varphi)$ and $\vartheta(Q, \varphi, \theta)$, and then we will know T (and could simply compute $TSP = \exp(T)$). Uncertainty about these means that the functions $T^0(Q, \varphi)$ and $\vartheta(Q, \varphi, \theta)$ are themselves random quantities and so is T.

In order to derive the A-distribution of T and hence the value $t_{1-\gamma}$ (which is the exponential of the lower $(1-\gamma)$-quantile of that distribution) we need the A-distributions of Q, φ and θ, and in particular we need to identify the unknown features of those distributions that I have denoted by δ. As already explained, EVS 2010 treats the A-distribution of φ as a known discrete probability distribution, with no uncertain features. In contrast, it effectively treats Q as having a completely unknown A-distribution. Finally, the distribution of θ is assumed to be completely known — EVS 2010 says that θ is made up of a number of components, all of which are assumed to be normal with zero means and known variances, although the methodology does not depend on these details. I will return to the assumption of known distributions for θ when addressing a similar assumption for some E-distributions.

8 Reducing δ

Even with A-distributions assumed completely known for φ and θ, there remains the problem that the A-distribution of the ripples Q is complex and cannot be assumed to follow any structural form. So the set of unknown 'features' of this distribution that I have denoted by δ is effectively infinite. Before addressing the kind of data that are available and the specific solution adopted in EVS 2010, we can identify a range of possible approaches.

1. We could accept that the A-distribution of Q needs a nonparametric treatment (which in statistics is a term used to indicate that nothing is assumed

about the distribution). We could attempt to obtain data that provide samples from this distribution, in order to learn about the distribution as a whole.

2. We could note that we actually don't need to learn about the high-dimensional quantity δ but in fact need only to learn about $t_{1-\gamma}(\delta)$, which is a scalar quantity. We could attempt to obtain data that provide evidence directly for $t_{1-\gamma}(\delta)$.

3. As a hybrid between those two extremes, we could represent $t_{1-\gamma}$ in terms of a minimal number of features, say $\delta^{\min}$, and attempt to obtain data to learn about $t_{1-\gamma}(\delta^{\min})$.

I will return to approach 1 in Section 14. Approach 2 is appealing but obtaining direct evidence about a quantile of the A-distribution of T is problematic; no way of doing that occurs to me. EVS 2010 adopts approach 3. It expresses $t_{1-\gamma}$ as

$$t_{1-\gamma} = \mu_T + r_{1-\gamma}\sigma_T \ , \tag{6}$$

where μ_T and σ_T are the mean and standard deviation of the A-distribution of T and $r_{1-\gamma}$ is in effect defined as $(t_{1-\gamma} - \mu_T)/\sigma_T$. This reduces δ to $\delta^{\min} = (\mu_T, \sigma_T, r_{1-\gamma})$. Data are then used to learn about these. However, the data are used in two different ways. Statistical estimators are developed for μ_T and σ_T and their E-distributions are carefully derived (albeit asymptotically and approximately). The tolerance limit is based on these E-distributions. In contrast, although $r_{1-\gamma}$ is estimated from data the E-distribution of this estimator is not considered and is not part of the tolerance limit calculation. Uncertainty about $r_{1-\gamma}$ is therefore not accounted for. Furthermore, it is apparently estimated by computing the analogue of $(t_{1-\gamma} - \mu_T)/\sigma_T$ in the distribution of a quantity V that can be considered a noisy version of T. The difference between the two distributions will induce an error in $r_{1-\gamma}$ that could be important because the distribution of V is likely to be less skewed than that of T.

Remark 8 *The estimation of $r_{1-\gamma}$ in equation (6) gives cause for concern because it employs a surrogate device that in this case could lead to a bias, and because uncertainty in the estimate is not accounted for in the theory.*

This is potentially an important consideration. Even a relatively small error or bias in the estimation of $r_{1-\gamma}$ could influence the computed trip setpoint appreciably.

9 The E-equation

The second key equation in the EVS 2010 approach is (EVS-58), which we write here as

$$U_j(\varphi) = T^0(Q_j, \varphi) + \tau(Q_j, \varphi, \varepsilon_j) \ . \tag{7}$$

I will refer to this as the *E-equation* because it concerns historic data that are used to learn about (μ_T, σ_T).

The subscript j indexes data obtained from different historic snapshots of the reactor's operation. Each instance will be associated with its own ripples Q_j. In principle it would also be associated with its own flux shape φ_j, although the historic data are obtained in normal operating conditions, so that the flux shape should in fact be flat. EVS 2010 treats (7) as conditional on an assumed value of φ. On the left hand side, $U_j(\varphi)$ is a computation of the logarithm of TSP computed from the j-th data snapshot, but with a different set of substitutions:

(a') we estimate Q_j using the SORO code and compute COP_i and FOP_i from the assumed φ;

(b') we replace $FX_{CH,k}$ by measured fluxes from the data;

(c') we replace RP by its estimate RP^{ind}; and

(d') we replace CCP_i by its value computed using the assumed φ and the reference hydraulic conditions.

Then $U_j(\varphi)$ is regarded as an estimate of $T^0(Q_j, \varphi)$. The error term $\tau(Q_j, \varphi, \varepsilon_j)$ depends on Q_j and φ and on the random deviations ε_j in the above substitutions. That is, ε_j represents errors in the SORO estimate of Q_j and in the RFSP computation of φ for the assumed scenario, errors in the measured fluxes used for $FX_{CH,k}$ and RP^{ind}, and the error in CCP_i due to deviation of thermal hydraulic conditions from the reference values.

The evidential uncertainties are therefore represented by Q_j, φ and ε_j, and for the second part of the tolerance interval construction we need their E-distributions. It is assumed that each Q_j follows the same distribution as the Q in the future actual instance, and so its E-distribution is the same as the A-distribution of Q.

Remark 9 *The assumption that the Q_js follow the same distribution as Q may not hold if there are reasons for the ripples distribution to change over time. For instance, reactor management practices, and in particular the refuelling practices, may change. Also, as reactors age we may expect the ripples distribution to change slowly.*

The E-distribution of ε_j is assumed to be completely known, for instance with each component having a normal distribution with zero mean and known variance.

Remark 10 *The error random variables θ and ε that enter the A-equation and the E-equation respectively are assumed to have known distributions. In response to my questions about this, the OPG/BP representatives have asserted that the distributions are based on extensive evidence. However, this is not entirely based on comparison of estimates with known true values, since that is not always possible. EVS 2010 uses a surrogate approach to this which is considered in the next subsection.*

The E-distribution of φ is the same as the assumed A-distribution. Independence between Q_j and ε_j across j is another reasonable assumption as long as there is a sufficient time lapse between adjacent snapshots. This limits the frequency with which snapshots can be taken for use in EVS 2010, and so limits the available quantity of data.

10 The data

A key part of the modelling in EVS 2010 is that the only unknown feature of the A-distributions is the distribution of Q, and that the E-distributions depend only on the same distribution. There are no additional nuisance parameters. This is achieved primarily by the assumption of known distributions for θ and ε_i, but also through the E-equation being made conditional on a φ which is given the same distribution as in the future actual instance.

The significance of conditioning on φ is that $U_j(\varphi)$ can be evaluated from the j-th dataset for any arbitrary φ. The data are effectively being used to provide information about the A-distribution of Q, which varies in historic normal operation of the reactor in the same way as in a future instance.

Although $U_j(\varphi)$ could be computed for many values of φ and each j, the resulting data would not be independent and would be complex to use. Computing $U_j(\varphi)$ with a single φ randomly selected for each j would yield independent data but would add noise. The solution adopted in EVS 2010 is to compute $U_j(\varphi)$ for all the discrete φ values (or all in some category) and to average with respect to the assumed distribution over those φ values. I will denote the result by $\bar{U}_j$. These are independent random variables.

At this point it is necessary to consider the link from the $\bar{U}_j$s to T that will be used to construct the tolerance limit. From the A-equation and the E-equation we can write $U_j(\varphi)$ as T adjusted by two error terms:

$$U_j(\varphi) = T + \vartheta(Q, \varphi, \theta) - \tau(Q_j, \varphi, \varepsilon_j) \ .$$

Thus, after averaging over the various uncertain quantities, we can consider each $\bar{U}_j$ as an estimate of T. However, note that although the individual components of θ and ε_j may have nice symmetric distributions with zero means and known variances this will not in general be true of the derived error terms ϑ and τ. This is recognised in EVS 2010, which allows the means of these derived errors to be non-zero and their variances to be non-constant. However, these means and variances are assumed to be known. They can in principle be derived from the distributions of θ and ε_j, although again the surrogate approach is used.

Remark 11 *The surrogate approach for evaluating the means and variances of error terms is described in the EVS 2010 section EVS-6, where its accuracy is demonstrated in some examples. Nevertheless, estimating them from a finite dataset implies some uncertainty as to their values, and no matter how much data is used for these assessments they will necessarily be subject to some imprecision due to the unavoidable use of surrogates.*

EVS 2010 finds it convenient to adjust the $\bar{U}_j$s to remove the non-zero means of the error terms, in order to obtain unbiased estimation of T. The mean and variance of $\vartheta(Q, \varphi, \theta)$ conditional on Q (i.e. taking mean and variance with respect to the A-distributions of both φ and θ) are denoted respectively by $X^{\vartheta}(Q)$ and $Y^{\vartheta}(Q)$. Similarly the mean and variance of $\tau(Q_j, \varphi, \varepsilon_j)$ conditional on Q_j (taking mean and variance with respect to the A-distribution of φ and the E-distribution of ε_j) are denoted by $X^{\tau}(Q_j)$ and $Y^{\tau}(Q_j)$. Then we define

$$V_j = \bar{U}_j + X^{\vartheta}(Q_j) - X^{\tau}(Q_j) \ , \tag{8}$$

where the SORO estimate of Q_j for the j-th dataset is used to evaluate all three terms on the right hand side. The EVS 2010 document then shows that the expectation of each V_j is μ_T and derives a formula (at the bottom of page 33 of the EVS 2010 document) for the variance σ_V^2 of V_j in terms of σ_T^2 and several other quantities.

Remark 12 *The uncertainty/inaccuracy in the SORO estimate of Q_j is accounted for as part of ε_j and forms a part of the variance of $\bar{U}_j$. However, the formula for the variance of V_j does not also include a similar allowance for the use of the estimate in place of the true Q_j when evaluating the remaining terms $X^{\vartheta}(Q_j) - X^{\tau}(Q_j)$ on the right hand side of equation (8).*

11 The EVS 2010 tolerance limit

Using the V_js as the data, EVS 2010 chooses to set a tolerance limit of the form

$$T^* = \bar{V} + \lambda S_V \ , \tag{9}$$

where $\bar{V}$ and S_V are the sample mean and standard deviation of the V_js. Now it is necessary to derive the E-distribution of T^* in order to solve the tolerance limit equation (4) to obtain the value of λ.

At this point, EVS 2010 makes use of some sophisticated statistical theory which shows that if the number n of historic data snapshots is sufficiently large the E-distribution of T^* is approximately normal with mean $\mu_V + \lambda\sigma_V = \mu_T + \lambda\sigma_V$ and with a variance that depends in a complex way on μ_V, σ_V^2 and on the third and fourth moments of V. The use of this theory is an important part of the methodology, because it makes no assumptions about the shape of the distribution of V_j (other than the existence of fourth-order moments). Because of the complex nature of the basic formula (1) for the ideal TSP, and in particular because of the presence of min and max operations, the distribution of V_j can be expected to be skew and possibly far from normal. The methodology explicitly recognises this (and this recognition gives rise to the original use of 'EVS' to stand for 'extreme value statistics').

Remark 13 *The theory rests on the asymptotic normality of $\bar{V}$ and S_V, It thereby introduces a further approximation, that their joint distribution is exactly normal and that the asymptotic first and second-order moments equate to their* *actual moments. The quality of this approximation clearly rests on the size of the dataset.*

The resulting tolerance limit is a complex formula in which various quantities appear that must be estimated, notably the various constituents of σ_V^2 which have not been set out explicitly here, the first four moments of V_j (which are estimated using the sample moments) and a number of other quantities that have been identified in preceding sections of this review.

Remark 14 *The EVS 2010 tolerance limit is an ingenious solution to a highly complex and challenging problem. The use of the tolerance limit framework for setting a TSP is original in this field and very appropriate; the required separation of A- and E-uncertainties is carried out quite carefully and rigorously; the recognition that the derived error terms in the A- and E-equations may have non-zero means and non-constant variances, together with the recognition that the E-distribution of T^* may be far from normal, are excellent features of the approach. Numerous simplifications and assumptions must almost inevitably be made to tackle such a problem. Most of these are explicitly recognised in the EVS 2010 document and carefully justified, but it must be appreciated that the accuracy of the derived tolerance limit depends on the accuracy of all these assumptions and simplifications.*

Part III
Evaluation

My evaluation of the EVS 2010 methodology is first that it is mathematically and statistically sound, but on the question of whether it is fit for purpose I have a number of reservations. These reservations are set out in the next four sections.

- *Accounting for uncertainty.* As noted in Remark 14, the validity of any tolerance limit computed using EVS 2010 relies on the validity of the assumptions and approximations that underlie it. Approximations are unavoidable in challenging problems such as this, but they lead to additional uncertainty. I have reservations over the extent to which the relevant sources of uncertainty are accounted for in the final TSP.

- *Integrating out the flux shape.* A specific objective of EVS 2010 is to meet a criticism of an earlier approach, which suggested integrating out the flux shape. EVS 2010 does this in a technically sound and ingenious way. My reservation, though, is partly over the need to account for uncertainty in the specification of probabilities or weights, and partly over whether this integration is actually appropriate.

- *A more direct solution.* The EVS 2010 solution is complex, particularly where it integrates out the flux shape. I am concerned over the robustness of the resulting equations, and wonder whether a more direct simulation approach would be worth considering.

- *Reducing E-uncertainty.* Table 2 of the EVS 2010 document suggests that decreasing E-uncertainty regarding the accuracy of the physics codes used may lead to a decreased TSP. I have described this in correspondence with OPG/BP as paradoxical because it goes against my intuition that decreased uncertainty should allow a higher setpoint. The fact that this has not been resolved yet in those discussions is another source of concern.

My full evaluation is presented in the final Conclusions and Recommendations sections.

12 Accounting for uncertainty

One motivation for a tolerance limit is to account for uncertainty in the A-distribution, as represented in this review by the uncertain parameters δ. If there were no such uncertainty, we would not need the tolerance limit equation (4) but could simply compute the required quantile $t_{1-\gamma}$ defined in (3). As mentioned in Section 4, if there are additional uncertain parameters, known as nuisance parameters, then the tolerance limit formula (4) has to hold for all values of those nuisance parameters, which is often difficult to achieve. EVS 2010 makes a number of assumptions and approximations which mean that there are no nuisance parameters and δ is reduced to a minimal parameter set $\delta^{\min} = (\mu_T, \sigma_T^2, r_{1-\gamma})$. Various other assumptions, simplifications or approximations have been noted in this review. I will refer to them collectively as *compromises*. Every one of these compromises implies that the final result is approximate to some degree and is a source of uncertainty that should be accounted for. The following is a list of all such sources of additional uncertainty.

1. Remark 2 points out that the basic equation (1) for the ideal TSP implies an assumption that fuel channel powers and fluxes at detector locations in safety channels will all increase in direct proportion to overall reactor power.

2. Remarks 5 and 6 highlight the assumption that the possible flux shapes form a discrete set and that the finite set which has been identified for a given reactor is complete.

3. Remark 7 concerns the probabilities (also referred to in EVS 2010 as weights) for the different flux shapes. There is undoubtedly uncertainty concerning these, particularly when a uniform distribution is assumed across a category of abnormal shapes.

4. Remark 8 identifies the estimation of $r_{1-\gamma}$ as another unaccounted source of uncertainty, and as possibly subject to bias. This is important because even quite small errors in $r_{1-\gamma}$ are likely to influence the final TSP appreciably.

5. Remark 9 concerns the possibility that the distribution of ripples may change over time, and so the assumption that the distribution of ripples in historic snapshot data is the same as in a future LOR instance may not hold.

6. Remarks 10 and 11 are concerned with the assumptions that the distributions of θ and ε_j, in particular their means and variances, together with various other quantities involved in σ_V^2 are known. In practice these are estimated, and estimation error arises not only from the finite quantity of data or simulations used but more importantly from the use of a surrogate methodology.

7. Remark 12 points out that uncertainty due to SORO estimation of ripples Q_j is accounted for in $\bar{U}_j$ but not in $X^\vartheta(Q_j)$ and $X^\tau(Q_j)$.

8. Remark 13 notes that the asymptotic theory used in EVS 2010 will not hold exactly in a finite sample, and so this is another approximation.

EVS 2010 accounts for the uncertainty implicit in using the sample mean and standard deviation of the V_is to estimate the population mean and standard deviation μ_V and σ_V, through the use of their E-distributions. However, all of the above compromises imply additional sources of uncertainty that are not currently accounted for in the EVS 2010 methodology. In principle, they might all be expected to lead to reduced trip setpoints in order to fully achieve the β confidence requirement in (4). The size of such impacts is a key factor in deciding whether EVS 2010 is fit for purpose.

13 Integrating out the flux shape

Uncertainty about the flux shape has been identified in the preceding section as an important issue. If the probability distribution of φ had been estimated from a sample, then the usual multinomial distribution (or the Dirichlet posterior in a Bayesian analysis) could be used to quantify uncertainty about the probabilities. However, it seems from correspondence with OPG/BP that at least within some categories the probabilities/weights are set to be uniform. If this is done simply because of lack of evidence then it could be considered to arise from an exchangeable Dirichlet prior distribution — that is, a Dirichlet distribution with parameter vector $(q, q, \ldots, q)$ in which all the k elements are the same, where k is the number of possible flux shapes. The value of q determines the strength of the prior information. Setting $q = 1$ is a common choice, but $q = 1/k$ would be more cautious. Some such choice could be argued for as a default characterisation of uncertainty about φ. Notice that such a distribution would entail considerable uncertainty (which is of course appropriate when there is essentially no data), and this would have a substantial impact on the TSP.

As noted in Section 6, a uniform distribution may also be a deliberate decision, defining the A-population of future events to be one in which the flux shapes are equally probable. For the moment, suppose that the flux shape is the only A-uncertainty. Then the quantile equation (3) will in general set $t_{1-\gamma}$ so that there is a $100\gamma\%$ chance of the NOP trip operating in time, for a future instance in which just the flux shape is uncertain. However, setting all the flux shape probabilities equal (to $1/k$) gives this equation an alternative interpretation — $t_{1-\gamma}$ is chosen so that the NOP trip will operate in time for $100\gamma\%$ of the possible flux shapes.

In general, we can recognise at least three cases:

- The NOP trip should operate in time with $100\gamma\%$ probability for a random future instance in which flux shape is also random (within the desired category). In this case, uncertainty about the probability distribution of flux shape should be accounted for, as discussed above.

- The NOP trip should operate in time with $100\gamma\%$ probability for a random future instance defined to have equal flux shape probabilities. The $100\gamma\%$ is now an average over the possible flux shapes (in the desired category, with equal weights). In this case, the second interpretation above applies, but the presence of other A-uncertainties means that it is a little more complex and now refers to an average probability.

- The NOP trip should operate in time with at least $100\gamma\%$ probability for a random future instance and for all possible flux shapes (in the desired category). This is a worst-case approach to flux shape.

The presentation in EVS 2010 suggests that the first of these cases applies, so uncertainty about the distribution of φ should be quantified. Uncertainty about the distribution is not relevant in the second case, and integrating out with a uniform distribution is correct. In the third case, no distribution is needed and it is not appropriate to integrate out φ.

Which of these cases (or indeed some other case) applies is a deliberate choice about what neutron overpower protection is wanted. However, it may be of interest to note that if we take the first case with an exchangeable Dirichlet distribution, setting q to be large leads to the second case (equal weights with no uncertainty), whereas a small q will approximate to the third case (protect against worst flux shape).

14 A more direct solution

The EVS 2010 solution involves rather complex equations. In particular, the value of λ in equation (9) is obtained from a calculation involving a number of quantities which must be estimated, including third and fourth moments of the distribution of the V_js. Even if the individual components are estimated with only small errors, the number of them and the fact that some are subtracted from others may lead to a lack of robustness in this computation.

Part of the complexity arises from the fact that the data are based on $\bar{U}(Q_j)$ which because φ has been integrated out will show less variability than T. Thus, in general we can expect σ_V to be less than σ_T and the calculation involves compensating for this through one of the estimated terms. Part also arises from the fact that the data are being used essentially just to learn about the ripples distribution, yet the A- and E-equations are linked not directly through the ripples but more indirectly through T^0. (Further loss of efficiency arises from the fact that $\bar{V}$ and S_V will not be formally sufficient for μ_V and σ_V, although this is likely to be a very small effect.)

It may be worth considering a technically simpler and more direct approach. In general, a tolerance limit can be computed by a two-stage Monte Carlo approach.

1. Draw a random set of values of the unknown parameters δ in the A-distributions, from their E-distributions.
2. Inner loop:
   (a) Sample all the A-distributions using the current sampled values of δ from the outer loop.
   (b) Compute TSP.

   (c) Repeat the above two steps a large number of times to obtain a large sample of TSP values.

   (d) Set $t_{1-\gamma}$ to be the value such that it is exceeded by $100\gamma\%$ of sampled TSP values.

3. Repeat all the above to obtain a large sample of $t_{1-\gamma}$ values.

4. Set T^* to be the value such that it is exceeded by $100\beta\%$ of sampled $t_{1-\gamma}$ values.

In the context of the NOP trip setpoint problem, step 1 involves sampling from the distributions of all the uncertain parameters that determine the A-distributions of Q, φ and θ. EVS 2010 addresses uncertainty in the A-distribution of Q but assumes that the distributions of φ and θ are completely known. In this more direct solution, it becomes possible to treat those distributions as uncertain. Uncertainty in the distribution of θ can be accommodated by assigning an E-distribution to the variances of the components of θ, and possibly also distributions to their means (centred on zero but not assuming zero means). Uncertainty in the distribution of φ can be accommodated by sampling from a Dirichlet distribution or any other appropriate representation of uncertainty in the weights/probabilities.

Uncertainty in the A-distribution of Q needs to be handled differently from EVS 2010 because in essence this Monte Carlo solution is Bayesian. In effect we employ approach 1 of Section 8. First suppose that there is no error in the SORO computations, so that from each historic snapshot we have an observed Q_i. From this sample we can derive a posterior distribution for the cdf of Q — the simplest approach would be using a weak Dirichlet process prior that results in Rubin's Bayesian bootstrap. The regular bootstrap would be a simple alternative way of sampling from the posterior. With normally distributed model inadequacy error in SORO we would obtain instead a posterior that is a Dirichlet mixture of normals and also simple to sample from. (This is the only part of ε_j that is needed in this approach.)

Such a solution would be computationally intensive but I think it would be feasible. It would be able to address the extra uncertainties in points 3, 4, 6 and 7 of Section 12.

15 Reducing E-uncertainty

Table 2 of the EVS 2010 document suggests that the methodology behaves in a way that is counter-intuitive, at least for me. I have called this a paradox in my discussions with OPG/BP and despite their objections to this term I stick by it (see Appendix C). I mean simply that there is an apparent contradiction, between the claimed behaviour of the EVS 2010 trip setpoint and my intuition.

In Table 2, the TSP under nominal errors (i.e. error standard deviations for the components of ε as they are known or actually estimated to be), with random ripples and flux shapes, is 128.4. If the uncertainty in some of the physics codes is reduced, referred to in Table 2 as 'reduced' errors, then the TSP falls to 126.0. The paradox is this. My intuition is that if we decrease E-uncertainty then there should be less uncertainty about $t_{1-\gamma}$ and consequently it should be possible to have an increased TSP. Yet Table 2 shows the opposite behaviour — the TSP is decreased.

There is some extensive discussion of this issue in Appendix C which I will not repeat here. One key point is that in a recent meeting and correspondence OPG/BP have said that the 'reduced errors' in Table 2 refer specifically to reducing the *estimates* of those error variances, effectively deliberately under-estimating them from the available data. It is still somewhat counter-intuitive that estimating smaller error variances leads to reduced trip setpoints, but this is no longer so paradoxical. Indeed, both I and AMEC NSS have sketched alternative explanations, but which explanation (or both or neither) underlies the behaviour is still not clear. Furthermore, some other results presented by OPG/BP suggest that a similar paradoxical behaviour may exist with some A-uncertainties.

Despite several rounds of discussion, these matters are not fully resolved and I continue to refer to the behaviour as paradoxical. It will be important to resolve these in the benchmarking phase of this project.

16 Conclusions

The EVS 2010 methodology is conceptually and mathematically complex. I have found the document to be a challenging read, not least because there are many things that are not spelt out which I think would have benefited from much more explanation. However, I have found the representatives of CNSC, OPG, BP and AMEC NSS to be willing to assist me and their explanations have been almost invariably valuable. Appendix C documents partially my progress towards understanding.

I have found the mathematical and statistical analysis to be correct. The use of a tolerance limit approach in EVS 2010 is sound and entirely appropriate. The separation between uncertainties in an actual future instance when the NOP trip might be activated (A-uncertainties) and the uncertainties in data used to learn about the unknown parameters of the A-distributions (E-uncertainties) is fundamental to the tolerance limit approach and is carefully enforced in EVS 2010. I find that the terminology of aleatory and epistemic uncertainty used for this distinction in EVS 2010 does not agree with the way those words are generally used, which is why I refer instead to A- and E-uncertainty. The letters A and E stand for 'actual' and 'evidential' but have also been chosen for their mnemonic value if they are read instead as 'aleatory' and 'epistemic'.

In Section 1, four questions were identified as defining the scope of this review. The first was whether the EVS 2010 theory is mathematically and statistically correct, and the above paragraph makes it clear that my response to that question is positive. However, the second question is whether the method is fit for the purpose of computing NOP trip setpoints. On the question of fitness for purpose I find some issues of concern. In particular, EVS 2010 makes a number of simplifications, approximations and assumptions, generically called compromises, and some of these may materially affect the computed trip setpoint.

1. The methodology rests on equation (1) which expresses the ideal trip setpoint under perfect knowledge of the reactor conditions, and assumes that this is an accurate formulation of the NOP trip setpoint problem. It also embodies an assumption that in an LOR event the fuel channel powers and the fluxes at detector sites in the safety channels will all increase in strict proportion to the overall reactor power. This aspect of the superposition principle seems to be an accepted position in the industry, but I feel it should be explicitly acknowledged as an assumption.

2. The γ/β tolerance limit rests on frequentist statistical theory, rather than Bayesian theory, and this affects its interpretation. The tolerance limit provides $100\beta\%$ confidence that the NOP trip will activate sufficiently early in at least $100\gamma\%$ of future actual instances. The frequentist expression '$100\beta\%$ confidence' should be interpreted as saying that if trip setpoints were computed using the tolerance limit method on a large number of random repetitions of the dataset used, then $100\beta\%$ of those trip setpoints will have the required property (viz. that the NOP trip will activate sufficiently early in at least $100\gamma\%$ of future actual instances). A Bayesian analogue would replace the word 'confidence' with 'probability', and would thereby have the more useful interpretation that the trip setpoint computed from the one actual dataset has $100\beta\%$ probability of having the required property. I do not mean to imply that the fact that EVS 2010 uses the 'confidence' statement is in itself a concern. My concern is that the industry needs to be aware of the difference. The question of the meanings of γ and β, although subtle, is important because it is fundamental to the nature of the protection provided by the trip setpoint.

3. EVS 2010 uses data from historic snapshots of reactor operation as an intrinsic part of the tolerance interval. There is an assumption that the past is an accurate guide to future uses of the trip setpoint, in the sense that the ripples distribution remains constant over time. This should be recognised explicitly because in practice it may not hold, due to changes in operating practices or ageing.

4. 'Integrating out the flux shape' is a feature of the EVS 2010 solution. Formally the flux shape, denoted by φ, is one part of the uncertain or random reactor conditions in an actual future instance. EVS 2010 accounts for this uncertainty through a probability distribution for φ which is part of the A-uncertainty specification and is also used to 'integrate out' φ in the data calculations. The defnition and nature of this distribution are intimately linked to the definition of the future instance for which protection is sought, and thereby to the value of γ. Several issues concern the nature of the flux shape distribution.

   - In EVS 2010 φ takes a finite set of possible values and its distribution is discrete. This is in fact a simplification of what should in reality be continuous.
   - In addition it is possible that the discrete set employed in EVS 2010 may not contain all possible flux shapes. Even if excluded shapes are deemed very rare, if they would lead to very low ideal trip setpoints their exclusion will lead to a computed trip setpoint that is too high.
   - Whether one treats φ as random at all depends on a decision regarding the kind of neutron overpower protection that is desired. But if it is (which is the position taken in EVS 2010) then there will be appreciable uncertainty concerning its distribution, and such uncertainty is not accounted for in EVS 2010. This is an issue of particular concern.

   The first two of these points may turn out to have only minor impact on the final trip setpoint. In contrast, I believe the question of uncertainty in the distribution of φ is important and should be addressed.

5. Several of my concerns focus on other elements of uncertainty that I believe should be attached to the final TSP but which are not accounted for in EVS 2010.

   - Estimation errors will arise in specifying the distributions of the error variables denoted in EVS 2010 by θ and ε and in various other terms appearing in the TSP solution through the formula for σ_V^2 (the variance of each data item V_j), partly because of the limitations of finite sampling and partly from the use of a surrogate approach. These terms are assumed known in the EVS 2010 theory and uncertainty in them is not accounted for.
   - The EVS 2010 tolerance limit is constructed using a representation of the quantile trip setpoint $t_{1-\gamma}$ in terms of quantities μ_T, σ_T and $r_{1-\gamma}$. The first two of these are formally estimated and uncertainty in those estimates is accounted for through their E-distributions. However, uncertainty in the estimate of $r_{1-\gamma}$ is not accounted for, and nor is the fact that the estimate derives from a sample of V values whose distribution is different from that of T. The computed trip setpoint is likely to be sensitive to $r_{1-\gamma}$, and so this is an issue of particular concern.
   - Uncertainty regarding the SORO estimates of ripples factors is only partially accounted for in the formula for σ_V^2.
   - EVS 2010 uses asymptotic theory to represent the sample mean and standard deviation of the data V_j as having normal distributions. The number of data points is finite, so this is necessarily an approximatioin.

The first, third and fourth of these may turn out to have only minor impact on the final trip setpoint. However, the estimation of $r_{1-\gamma}$ is intrinsically more important and the attendant additional uncertainty and possible bias should be addressed.

6. The EVS 2010 tolerance limit solution is a sophisticated and ingenious piece of statistical theory. It also is quite complex to implement (and implementation will require many details and explanations that are not in the report itself). Two issues around this are worth noting.

    - It may be that further approximations are actually used in implementation without being noted within the EVS 2010 document.
    - I am concerned at the complexity of the final solution which it seems to me might not be robust to estimation errors. The complexity arises partly from the integrating out of φ in the definition of the data items V_j and partly from the fact that data whose role is primarily to learn about the ripples distribution are linked more indirectly to the ideal TSP.

7. Some claims made in the EVS 2010 document and in subsequent discussions indicate paradoxical behaviour of the EVS 2010 computed trip setpoint when uncertainties, or estimates of uncertainties, are reduced. This remains another issue of particular concern until the apparent contradictions are resolved.

Having listed all the various compromises that EVS 2010 makes, we must recognise that the NOP trip setpoint problem is difficult and it can hardly be expected to yield to a perfect solution. Any method for deriving NOP trip setpoints must surely make some compromises, and the fact that they are made in EVS 2010 does not in itself imply that the methodology is not fit for purpose. Fitness for purpose is a judgement about whether despite all of these compromises the method will give solutions that are sufficiently accurate.

Remark 15 *Fitness for purpose is a judgement about whether a method is sufficiently accurate, recognising that perfection is not a practical possibility. There are two aspects to this judgement. One is to assess how accurate the method is. It is usually not possible to know precisely how accurate it is because we don't have a perfect solution against which to compare it, so this requires a careful and informed judgment. The second judgement is whether that level of accuracy is acceptable.*

Ultimately, it seems to me, fitness for purpose of EVS 2010 is a matter for CNSC to decide. However, I have particular concerns about the flux shape distribution and the factor $r_{1-\gamma}$ and it is my own judgement that at least these should be explored further before EVS 2010 can be regarded as fit for purpose.

Returning to the four questions in Section 1 which define the scope of this review, the third question asks whether there are alternative approaches that should be considered and compared with EVS 2010. I feel that the tolerance limit approach, or its Bayesian analogue, is important, and therefore I do not regard earlier methods which do not use tolerance limits as competitors to EVS 2010. However, in Section 14 I have suggested a conceptually simpler and more direct alternative computation based on a nested Monte Carlo simulation. I believe that this approach should be feasible and is worthy of consideration in the future. The resulting trip setpoint would in this case have a Bayesian interpretation.

The fourth question in Section 1 asks whether there are other statistical techniques which might be used to improve EVS 2010. It is my assessment that EVS 2010 already makes use of key statistical theories in following through its particular approach to forming a tolerance limit. I have suggested in Section 6 that a bootstrap or Bayesian bootstrap technique could be used to incorporate some uncertainty in the flux shape distribution.

Finally, it may also be worth investigating the method known as emulation of complex computer codes. This may be useful to reduce the computational burden associated with the Monte Carlo simulations employed to estimate error variances and some other parameters. Since they are done only once for the NOP trip setpoint computation, and are presumably feasible within current computing resources, this may not be relevant. However, there is always a wish to use more and more complex and realistic computer simulation codes, and a major restraint on doing so is always the computing power needed. Emulation is a powerful tool for reducing the number of runs required to do analyses such as uncertainty and sensitivity analyses, typically requiring orders of magnitude fewer runs than Monte Carlo. As such it may have various roles within the industry, not just in the NOP trip setpoint problem.

17 Recommendations

On the basis of my conclusions presented in Section 16, I make the following recommendations.

Recommendation 1. The industry, and in particular the regulator (CNSC), should consider carefully the meanings of the two quantities γ and β in a tolerance interval, with reference to the NOP trip setpoint problem. The interpretation of γ requires careful specification of the random circumstances of the future instances in which the NOP trip is to operate, and in particular the probabilities or weights assigned to different flux shapes. In regard to β, the distinction between the 'confidence' interpretation of a frequentist tolerance interval (used in EVS 2010) and the probability interpretation of its Bayesian analogue is also important.

Recommendation 2. The impact of additional uncertainties that arise implicitly or explicitly from compromises in the EVS 2010 methodology should be assessed through benchmarking tests. Particular attention should be paid to the uncertainty in the flux shape distribution and in the estimate of the factor denoted in EVS 2010 by $r_{1-\gamma}$. Judgement regarding the fitness for purpose of EVS 2010 should be reserved until these tests have been conducted and evaluated.

Recommendation 3. The paradoxical reported behaviour of the EVS 2010 trip setpoint when uncertainties, or estimates of uncertainties, are reduced should be investigated and the paradox resolved, to determine whether it is the methodology (or its implementation) or common intuition which is at fault, or whether it is simply the result of a misunderstanding. Benchmarking tests have a useful role to play. This exercise is also important to carry out before making any fitness for purpose judgement on EVS 2010.

Recommendation 4. A more direct, Bayesian approach outlined in Section 14 of this review should be contrasted with EVS 2010 in benchmarking tests.

Although I make only these formal recommendations, it would be a mistake to focus only on these. I believe there are many more points in this review which, although not so important in my judgement that I make them formal recommendations, would nevertheless repay careful reading.

Part IV
Appendices

A Remarks

We list here the various remarks from the body of the report, that highlight key points.

Remark 16 *The NOP trip setpoint problem as formulated in EVS 2010 and as adopted here focuses on whether the channel power in any fuel channel exceeds its CCP. The ideal trip setpoint will ensure that the NOP trip operates in a slow LOR incident as soon as any channel power reaches its CCP. The validity of EVS 2010 depends in the very first instance on this being an accurate formulation of the NOP trip setpoint problem.*

Remark 17 *The definition of TSP in (1) or (EVS-5) implies an assumption that in an LOR event all channel powers increase in fixed proportion to the total power, and all flux detector readings also increase in fixed proportion to total power. This assumption has been acknowledged and defended by OPG and BP representatives as at least an accepted approximation to reality. The strength of the assumption is that it allows (1) to be used under any reactor conditions, even when overall reactor power is such that the reactor is nowhere near a dry-out condition, to deduce what TSP should theoretically apply under those conditions if reactor power were to increase due to slow loss of regulation. The EVS 2010 methodology relies on this assumption and to the extent that it is only an approximation this limits the validity of EVS 2010.*

Remark 18 *The tolerance limit approach adopted by EVS 2010 distinguishes between uncertainties relating to future actual reactor conditions (which I term A-uncertainties) from the uncertainties (E-uncertainties) relating to evidence used to learn about features of the A-uncertainty distributions. This is an important contribution to the NOP trip setpoint problem because it focuses on controlling the proportion of times (i.e. A-probability) that the NOP trip activates sufficiently early in all future actual instances.*

Remark 19 *The EVS 2010 objective is to compute a trip setpoint value T^* as a frequentist γ/β tolerance limit. It is important to recognise that the correct interpretation of this tolerance interval is that if we were able to repeat the computation of T^* many times using new sets of historic data then $100\beta\%$ of these T^* values would be below the value $t_{1-\gamma}$ at which the trip would successfully activate sufficiently early in $100\gamma\%$ of future actual instances. This does not imply that there is a $100\beta\%$ probability that the actual calculated T^* will be below $t_{1-\gamma}$. Such an interpretation can only be given for the analogous Bayesian limit.*

Remark 20 *EVS 2010 asserts that the set of possible flux shapes is discrete. The reality is that it is not strictly discrete, but treating it as such is not likely to lead to a too-high value for the NOP trip setpoint.*

Remark 21 *I have not received a convincing argument that the finite set of considered flux shapes is complete in the sense that there are not other scenarios which might be imagined. If others are possible and might lead to extremely low ideal trip setpoint values then, no matter how unlikely they are, to exclude them might result in over-prediction of the NOP trip setpoint.*

Remark 22 *The TSP is likely to be sensitive to the choice of probability distribution across flux shapes. In this context it is important to think carefully about the population of future events that determines the A-uncertainties, and to assign γ appropriately to that population. It is also important to recognise uncertainty about the flux shape distribution.*

Remark 23 *The estimation of $r_{1-\gamma}$ in equation (6) gives cause for concern because it employs a surrogate device that in this case could lead to a bias, and because uncertainty in the estimate is not accounted for in the theory.*

Remark 24 *The assumption that the Q_js follow the same distribution as Q may not hold if there are reasons for the ripples distribution to change over time. For instance, reactor management practices, and in particular the refuelling practices, may change. Also, as reactors age we may expect the ripples distribution to change slowly.*

Remark 25 *The error random variables θ and ε that enter the A-equation and the E-equation respectively are assumed to have known distributions. In response to my questions about this, the OPG/BP representatives have asserted that the distributions are based on extensive evidence. However, this is not entirely based on comparison of estimates with known true values, since that is not always possible. EVS 2010 uses a surrogate approach to this which is considered in the next subsection.*

Remark 26 *The surrogate approach for evaluating the means and variances of error terms is described in the EVS 2010 section EVS-6, where its accuracy is demonstrated in some examples. Nevertheless, estimating them from a finite dataset implies some uncertainty as to their values, and no matter how much data is used for these assessments they will necessarily be subject to some imprecision due to the unavoidable use of surrogates.*

Remark 27 *The uncertainty/inaccuracy in the SORO estimate of Q_j is accounted for as part of ε_j and forms a part of the variance of $\bar{U}_j$. However, the formula for the variance of V_j does not also include a similar allowance for the use of the estimate in place of the true Q_j when evaluating the remaining terms $X^{\vartheta}(Q_j) - X^{\tau}(Q_j)$ on the right hand side of equation (8).*

Remark 28 *The theory rests on the asymptotic normality of $\bar{V}$ and S_V, It thereby introduces a further approximation, that their joint distribution is exactly normal and that the asymptotic first and second-order moments equate to their actual moments. The quality of this approximation clearly rests on the size of the dataset.*

Remark 29 *The EVS 2010 tolerance limit is an ingenious solution to a highly complex and challenging problem. The use of the tolerance limit framework for setting a TSP is original in this field and very appropriate; the required separation of A- and E-uncertainties is carried out quite carefully and rigorously; the recognition that the derived error terms in the A- and E-equations may have non-zero means and non-constant variances, together with the recognition that the E-distribution of T^* may be far from normal, are excellent features of the approach. Numerous simplifications and assumptions must almost inevitably be made to tackle such a problem. Most of these are explicitly recognised in the EVS 2010 document and carefully justified, but it must be appreciated that the accuracy of the derived tolerance limit depends on the accuracy of all these compromises.*

Remark 30 *Fitness for purpose is a judgement about whether a method is sufficiently accurate, recognising that perfection is not a practical possibility. There are two aspects to this judgement. One is to assess how accurate the method is. It is usually not possible to know precisely how accurate it is because we don't have a perfect solution against which to compare it, so this requires a careful and informed judgment. The second judgement is whether that level of accuracy is acceptable.*

B Aleatory and epistemic uncertainties

In the main body of this report, I have used the terms A-uncertainties and E-uncertainties in preference to the terminology of aleatory and epistemic uncertainties that is used in EVS 2010. In this appendix I consider the origins of the distinction between “aleatory’ and ‘epistemic’ uncertainty with a view to explaining why I prefer A- and E-uncertainty in the context of NOP trip setpoint methodology.

B.1 Philosopy

The terms aleatory and epistemic are used in the philosophy of probability to mean the following.

- Aleatory (from the Latin, meaning a die, as in the famous phrase ‘*alea jacta est*’) refers to uncertainty due to intrinsically random events. Examples are the tossing of dice or the event of rain in a certain place on some date in the future.

- Epistemic (from the Greek, meaning relating to knowledge) refers to uncertainty due to lack of knowledge. Examples are the weight of President Obama or the event of rain in a certain place on this date ten years ago.

Other terms are used, sometimes for not quite equivalent concepts. Aleatory uncertainty is sometimes called irreducible, while epistemic uncertainty is said to be reducible. This is because where we have lack of knowledge we might in principle reduce that uncertainty by getting more information. For example, I could in principle resolve the question of rain ten years ago by consulting meteorological records or newspaper reports.

The distinction is often useful even though it is not completely clear cut. For instance, I could reduce my uncertainty about rain on a future date by consulting climatic data on the place in question, so even here there is an element of epistemic uncertainty. Yet no matter how much knowledge I assemble now, I would still be uncertain about that rainfall event so there will always be some residual randomness, i.e. aleatory uncertainty.

B.2 Statistics

In Statistics, the philosophical question primarily concerns whether the two different kinds of uncertainty require different forms of probability, and indeed whether epistemic uncertainties can properly be described by probabilities at all. In this respect, it is usual to interpret the terms in the following way.

- Aleatory quantities and events are in principle or conceptually repeatable, so that probabilities can be defined for them using the traditional long-run relative frequency formulation of probability. In particular, the data that are analysed in statistical methods are generally generated randomly and are assumed to have aleatory sampling distributions.

- Epistemic quantities and events are one-off and non-repeatable, so that they cannot be given probabilities according to the relative frequency formulation of probability. The unknown parameters in statistical models are generally considered to be epistemic.

The frequentist philosophy of statistics only acknowledges probability in the relative frequency sense. According to frequentist statistics, parameters are fixed but unknown, so they do not have probability distributions. Frequentist inferences, even when they appear to make probability statements about parameters do not do so, and indeed they cannot. Therefore, a 95% confidence interval does not say that there is a 95% probability that the (uncertain) parameter lies in the (fixed) interval. Instead it says there is a 95% chance that the (random) interval contains the (fixed) parameter. The probability statement is not about the parameter but about the interval.

Bayesian statistics is based instead on the personal (or subjective) formulation of probability, which applies to both aleatory and epistemic uncertainties. Bayesian methods therefore assign probability distributions to parameters. Indeed, Bayes' theorem works by combining a prior distribution for the parameters with the sampling distribution of the data in order to derive the posterior distribution of the parameters. Bayesian inferences are derived from the posterior distribution, and they do make probability statements about the parameters. For instance, a Bayesian 95% credible interval for a parameter does say that there is a 95% probability that the (uncertain) parameter lies in the (fixed) interval.

The frequentist/Bayesian debate has been long and sometimes bitter. Frequentists object primarily to the subjectivity of Bayesian methods. President Obama's weight is surely known by some people, others close to him may know his physique well enough to have quite good knowledge, while others who only see him on television have poorer information. This will be expressed in each having a different probability distribution for his weight. Bayesian inferences are technically subjective in this sense; they express the beliefs and knowledge of the individual analyst. However, advocates of the Bayesian view argue that (a) in reality subjectivity is minimised by good practice, and (b) to ignore prior information is just as bad science as to admit an element of expert judgement.

B.3 The nuclear context

The terms aleatory and epistemic are being seen increasingly often in scientific journals. Much of the impetus for this comes from the nuclear physics community and related fields. Since the 1970s or even earlier, the terms were being used in work of Sandia Labs and other US research institutions. And their usage has important similarities with the NOP trip setpoint problem.

A key feature of such problems is that the focus of interest is on some random future event (such as a slow LOR), and the essence of the problem is to control or predict that future event. Because it is random, it has aleatory uncertainty and we can postulate a probability distribution for it. In detail, the focus is on some feature F of this distribution, for instance the mean, the variance, the median or some other quantile (as in the NOP trip setpoint problem). However, we do not know F because we do not know the aleatory distribution of the future event. Specifically, this distribution has some unknown parameters X. And as unknown parameters, their uncertainty is epistemic.

This is typical of all statistical work, where uncertainty about future observations has an aleatory component via the sampling distribution of the observation and an epistemic component via the uncertainty in the parameters of that distribution. My example of rain on a future date is just like that. However, the type of problem that I describe above is characterised by the focus on some feature F of the aleatory future sampling distribution.

The work of Sandia and others typically assigns a probability distribution to X, and in doing so effectively adopts a Bayesian viewpoint. The problem is solved then by propagating this parameter uncertainty to obtain inferences about F. The terms aleatory and epistemic are correctly used here to describe the distribution of the future event and the distribution of X respectively. The distribution of X may be specified in a conventional Bayesian way as a posterior distribution based on some relevant data Z. It may also be specified by expert elicitation. In modern terminology, quantifying epistemic uncertainties with probabilities is described as Bayesian, even if Bayes' theorem is not used.

B.4 EVS 2010

EVS 2010 does not follow that route. Instead, it adopts an entirely frequentist approach. Data Z are used to derive a frequentist confidence interval for F. Note that in frequentist theory a confidence interval for a quantile is usually called a tolerance interval, and since the interval is one-sided in this case, we refer to it as a tolerance limit.

The reason why I do not think the terms aleatory and epistemic are used correctly in EVS 2010 is that the uncertainty about the data (Z) in a tolerance interval calculation is not epistemic but aleatory. In the EVS 2010 analysis the only unknown parameters (that I denote by δ in the main part of this report but which correspond to X in the above formulation) in the aleatory distribution for the future event are those that define the distribution of Q, but these are not explicitly identified in EVS 2010 and they certainly are not given a probability distribution. It is for this reason that I have preferred not to adopt the EVS 2010 terminology of aleatory and epistemic uncertainties.

It is nevertheless important in these problems always to identify two groups ot uncertainties. The first is the uncertainties that govern the future random event. These are correctly called aleatory in EVS 2010 (although the list of aleatory uncertainties in the document is incomplete because both Q and φ are also aleatory). I call them A-uncertainties. The second group is those that relate to the uncertainty in δ. In EVS 2010, these are the uncertainties that govern the random data, the U_is. I call these the E-uncertainties. EVS 2010 correctly keeps the two groups separate, so whether we call them aleatory/epistemic or A-/E- is academic.

RSP-0293

REPORT ON EVS 2010 GROUP A BENCHMARKING EXERCISE

Tony O'Hagan

22 February, 2012

EXECUTIVE SUMMARY

The Group A Benchmarking Exercise is part of the evaluation of the EVS 2010 methodology that has been developed by AMEC NSS and proposed for the NOP trip setpoint problem by OPG and Bruce Power. This report sets out the formulation and analysis of the Group A benchmark tests.

The Group A tests are set in a simplified scenario that does not retain some of the more complex features of the NOP problem but which nevertheless fits the statistical model which is solved in the EVS 2010 Report. Its simplicity allows rigorous exploration of the performance of EVS 2010 in tests for which true values are known and for which large numbers of datasets can be generated and analysed to reveal its statistical characteristics. In particular, the simplified scenario makes it possible to test whether EVS 2010 in practice behaves according to the tolerance limit criteria that it should theoretically satisfy. Using a simplified test-bed means that we cannot simply assume that behaviour in the tests will truly reflect behaviour in more complex applications. Nevertheless it is a useful guide.

A Bayesian comparator was developed in order to provide an alternative method to EVS 2010, because there was no pre-existing comparator that correctly separated the sources of uncertainty that the EVS 2010 Report calls aleatory and epistemic uncertainty, which is necessary to derive a tolerance limit.

The Group A benchmarking exercise comprised 22 tests in 6 suites. In each test, both EVS 2010 and the Bayesian comparator were used to analyse thousands of simulated sets of data of three different sample sizes. From each set of data, each method produced a computed value for the trip setpoint referred to as T*. The thousands of T* values computed for any given test, sample size and method were reduced to four summary output measures, known as *non-coverage, mean, SD* and *mean deficit*.

Various evaluation criteria were developed based on the four output measures. Evaluation Criteria 1 to 5 evaluated the performance of EVS 2010 in isolation against the behaviour that would be expected of a valid tolerance limit. Criteria 1 and 2 used respectively the non-coverage and mean deficit measures to assess formal compliance with the tolerance limit properties. Criteria 3 and 4 used the mean and SD measures to evaluate the face validity of EVS 2010 against forms of behaviour that it should logically have. Criterion 5 assessed its robustness to mis-specification by focusing on two of the test suites (Suites 4 and 6) in which assumptions were deliberately mis-specified. Evaluation Criterion 6 evaluated the performance of EVS 2010 against that of the Bayesian comparator with particular reference to how efficiently it used the available data.

The principal findings of the evaluation are summarised in the following two major recommendations.

MAJOR RECOMMENDATION 1: Subject to resolution of the face validity problem discussed in Major Comment 2 and Major Recommendation 2, and to any additional work deemed appropriate to address minor comments, EVS 2010 should be deemed to have passed the Group A Benchmarking Exercise.

MAJOR RECOMMENDATION 2: The developers of EVS 2010 should examine the causes of the paradoxical behaviour of EVS 2010 in Test Suite 3, with a view to justifying it or demonstrating that EVS 2010 will nevertheless behave acceptably in real applications for which it is proposed. Without satisfactory resolution of this problem, EVS 2010 should be deemed to have failed the Group A Benchmarking Exercise.

CONTENTS

INTRODUCTION AND OUTLINE

BACKGROUND

This report is prepared for the Canadian Nuclear Safety Commission (CNSC) under Contract 87055-10-1226 – R396.2: "Independent Verification and Benchmarking of Statistical Method and Mathematical Framework in OPG/BP 2010 EVS Methodology for Calculation of NOP Trip Setpoint".

The EVS methodology is set out in the document "A Genuine '95/95' Criterion for Computing NOP Trip Set-points Using EVS Methodology" by Paul Sermer and Fred Hoppe. That document is report number G0263/RP/008 from AMEC NSS Ltd., dated September 30, 2010, and will be referred to herein as the EVS Report. The methodology that it proposes will be referred to as EVS 2010. (Although the EVS Report was revised in 2011, the basic method remains as originally presented in the 2010 document.)

The contract identifies three substantive exercises as part of the "verification and benchmarking" of EVS 2010. The first is a technical evaluation of the correctness of the statistical and mathematical theory of EVS 2010. The second and third exercises are two groups of benchmarking tests to evaluate the performance of EVS 2010 in practice.

My report entitled "Review of the EVS 2010 Methodology", dated 6 August, 2011, presented my technical evaluation of the mathematical and statistical validity of the EVS 2010 methodology. It raised a number of concerns and made some recommendations for further action, including how the proposed benchmarking exercises might address some of those concerns. It will be referred to herein as my Review.

The Group A tests were to examine the performance of EVS 2010 in detail, using a simplified test scenario that should accord with the underlying statistical model assumed in the EVS 2010 theory, but without attempting to resemble the NOP trip setpoint problem. The Group B tests will be a more limited exploration based on realistic NOP trip setpoint scenarios.

This report presents the findings of the Group A benchmarking exercise.

CONDUCT OF THE GROUP A EXERCISE

The specification of the Group A tests was developed through a series of draft proposals presented by me. The first draft was presented to CNSC and revised following a teleconference with Dumitru Serghiuta on 8 August 2011. The second draft was circulated by CNSC to OPG and Bruce Power. Comments were received from AMEC NSS dated 15 August and a third draft presented on 6 September. Further comments were received from the industry, with response from me on 18 September.

It was clear that the industry still had serious concerns with the proposed benchmark test scenario and a teleconference was scheduled for 27 September 2011 between myself and representatives of CNSC, OPG, Bruce Power and AMEC NSS. The discussion was extremely useful and unearthed a misunderstanding on my part of the EVS 2010 method.

A fourth draft was presented on 13 October. This document, entitled "First Benchmarking Round: Proposal (Draft 4)", was adopted as the basic specification of the Group A benchmarking tests.

The proposal required parallel computations on a number of test cases using two different methods of generating computed trip setpoints. The EVS 2010 method was to be applied by AMEC NSS, while a method known as the Bayesian comparator was to be applied by CNSC. A document entitled "A Bayesian Method for the Benchmarking Exercise" was delivered on 16 October 2011, setting out the statistical basis of the Bayesian comparator. Computer code for implementing the two methods was given to me by AMEC NSS and CNSC, in order to confirm that they were correctly implementing the two methods and applying them correctly to the Group A tests.

Numerical outputs from the two methods were given to me on 4 November (Bayesian comparator) and 13 December (EVS 2010). I presented a first analysis of these results in a document entitled “Report on Group A Benchmark Tests“, that was circulated by CNSC to OPG and Bruce Power on 24 January, 2012. There followed an extensive and fruitful discussion of the results at a meeting in Ottawa on 31 January and 1 February.

This report sets out in detail the Group A benchmarking tests, the test outputs and my conclusions as informed by the Ottawa meeting. It subsumes and replaces the preceding three documents, namely “First Benchmarking Round: Proposal (Draft 4)”, “A Bayesian Method for the Benchmarking Exercise” and “Report on Group A Benchmark Tests“.

REPORT OUTLINE

The report is organised as follows.

Section “Recommendations of the 8 August 2011 report“ reviews the recommendations from my Review of the statistical and mathematical validity of EVS 2010, with particular reference to issues on which the Group A benchmarking tests are intended to cast some light.

Section “The Group A benchmark tests“ specifies the Group A tests, with the reasoning behind the inclusion of each suite of tests.

Section “The Bayesian comparator” describes the Bayesian comparator and discusses its strengths and weaknesses as part of the benchmarking exercise.

Section “Evaluation criteria” defines the test output measures and discusses how these were used in evaluating the performance of EVS 2010.

Section “Analysis of outputs” presents the numerical outputs of all the tests, and applies the evaluation criteria.

Section “Conclusions and recommendations” discusses the formal results of the evaluation criteria. My assessment of the implications of those results as regarding the validity and fitness for purpose of EVS 2010 are presented in a number of comments and recommendations.

RECOMMENDATIONS OF THE 8 AUGUST 2011 REPORT

My Review presented my technical evaluation of the mathematical and statistical validity of the EVS 2010 methodology as set out in the EVS Report. I found the theory to be mathematically and statistically correct, but had some reservations about how EVS 2010 would perform in practice. I expressed these concerns as relating to “fitness for purpose”.

In the Executive Summary of my Review I presented four recommendations, which are reproduced here.

Recommendation 1. The industry, and in particular the regulator (CNSC), should consider carefully the meanings of the two quantities γ and β in a tolerance interval, with reference to the NOP trip setpoint problem. The interpretation of γ requires careful specification of the random circumstances of the future instances in which the NOP trip is to operate, and in particular the probabilities or weights assigned to different flux shapes. In regard to β, the distinction between the `confidence' interpretation of a frequentist tolerance interval (used in EVS 2010) and the probability interpretation of its Bayesian analogue is also important.

Recommendation 2. The impact of additional uncertainties that arise implicitly or explicitly from compromises in the EVS 2010 methodology should be assessed through benchmarking tests. Particular attention should be paid to the uncertainty in the flux shape distribution and in the estimate of the factor denoted in EVS 2010 by $r_{1-\gamma}$. Judgement regarding the fitness for purpose of EVS 2010 should be reserved until these tests have been conducted and evaluated.

Recommendation 3. The paradoxical reported behaviour of the EVS 2010 trip setpoint when uncertainties, or estimates of uncertainties, are reduced should be investigated and the paradox resolved, to determine whether it is the methodology (or its implementation) or common intuition which is at fault, or whether it is simply the result of a misunderstanding. Benchmarking tests have a useful role to play. This exercise is also important to carry out before making any fitness for purpose judgement on EVS 2010.

Recommendation 4. A more direct, Bayesian approach outlined in Section 14 of this review should be contrasted with EVS 2010 in benchmarking tests.

Recommendations 2, 3 and 4 make explicit reference to using the benchmarking exercises to cast some light on areas of concern. The tests that make up the Group A benchmarking tests were formulated in part in response to these recommendations. The relationship of individual tests to the recommendations will be set out in the next section.

Recommendation 1 does not refer to benchmarking, and indeed relates more to the regulatory context underlying EVS 2010 than to the method itself. However, some of the discussion in the section “Evaluation criteria” may help to clarify the issues raised in this recommendation.

THE GROUP A BENCHMARK TESTS

Two groups of benchmarking tests were planned as part of the EVS 2010 evaluation project. The Group A test scenario should incorporate all the features of the NOP trip setpoint problem that are required to address issues in my Review, but can do so in a simplified and relatively abstract form. Group B is intended to create more realistic test cases.

The Group A benchmarking exercise will have two objectives. One is to check the performance of the method against truth. The purpose of this is to assess whether the claimed confidence level is achieved. If the computed setpoint is claimed to be a 95/95 tolerance limit, then we assess just what proportion of time the computed setpoint lies below the true lower 95% quantile value. In particular, we see how close this proportion is to the claimed 95%. This can only be done for artificial simulated problems, and the Group A benchmarking tests will do this in a simplified scenario.

In general, there may be many confidence limits which achieve the same stated confidence in repeated sampling. Whilst all valid 95% tolerance limits should on average lie below the true quantile 95% of the time, some might on average be much closer to that true quantile than others. In the same way that when faced with two unbiased estimators of a parameter we would prefer the one with the lower variance, so in the case of two valid 95% confidence limits we would prefer the one with the lower variance. The second objective of benchmarking involves assessing the performance of the EVS 2010 tolerance limit in this sense. In both cases the inference rule with the lower variance is preferred because it is getting more information value out of the data.

The comparison in this second kind of benchmarking is no longer against a known true values to assess the confidence claim, but is against alternative ways of constructing tolerance limits from the same data, to assess how well it makes use of those data. This aspect of the benchmarking must perforce be limited because there are no pre-existing competitors. EVS 2010 is the only proper tolerance limit method that has been proposed for the NOP trip setpoint problem. This is why the development of a Bayesian comparator was proposed in my Review, specifically for this benchmarking exercise.

ELEMENTS OF THE GROUP A TESTS

The Group A test suite has the following principal elements.

- A simplified scenario. The Group A scenario is a problem that retains all of the essential features on which the EVS 2010 theory rests, but which avoids addressing all the complexity of the full NOP problem.
- A base case in which baseline values are specified for all parameters of the simplified scenario. The base test forms a reference point against which we can compare other cases.
- Six suites of variations comprising another 21 specific tests. Each suite is designed to explore a particular issue or area of concern regarding the performance of EVS 2010. Each test is defined by varying the parameters in particular ways from the base case.
- In every test, both EVS 2010 and the Bayesian comparator were tested on a large number of random sets of data of three sample sizes, N = 20, 100, 500. Four principal outputs were recorded for each run.

These elements are developed in detail in the following subsections.

THE EVS 2010 STATISTICAL MODEL

The theory developed in the EVS Report rests on a statistical model having the following components.

1. There is a quantity T that we will refer to as the ideal trip setpoint, and which satisfies the equation
   $T = T^0 + \vartheta$.
   Note that in the real NOP trip setpoint problem as formulated in the EVS Report, T is the logarithm of the ideal trip setpoint. However, in the EVS Report all the statistical theory is developed on the log scale and only at the end is the actual trip setpoint computed as exp(T).
2. T^0 is a known function of two other quantities Q and φ that we refer to respectively as ripples and flux shape.
3. Q, φ and ϑ are random variables whose uncertainty is described here as A-uncertainty (and in the EVS Report as aleatory uncertainty). Because of this, T is also subject to A-uncertainty. The principal focus of attention in the statistical analysis is a low percentile of the A-distribution of T, which we refer to as $t_{1-\gamma}$.
4. The three variables Q, φ and ϑ are distinguished by having different degrees of knowledge about their A-distributions.
   - ϑ has a completely known A-distribution (including known variance).
   - The true A-distribution of the ripples Q is completely unknown.
   - The flux shape φ is known to take one of a finite number of possible shapes with known probabilities/weights, but we do not know those possible values of φ.
5. Because the A-distributions of Q and φ are unknown the A-distribution of T is unknown and so $t_{1-\gamma}$ is also unknown. In order to learn about the A-distributions of Q and φ, and so to learn about $t_{1-\gamma}$, we have some data. The data are, as in any statistical problem, random variables subject to observation errors. The uncertainty in these observations is referred to here as E-uncertainty (and in the EVS Report as epistemic uncertainty).
6. Separate data, with different characteristics, are available for Q and φ.
   - N random values are drawn independently from the A-distribution of Q. Each sampled value Q_j is observed as with error as S_j, according to the equation
     $S_j = Q_j + \varepsilon_j^{soro}$
   - An estimate Φ_k is available for each possible flux shape φ_k, satisfying the equation
     $\Phi_k = \varphi_k + \varepsilon_k^{rfsp}$
   - The errors ε_j^{soro} and ε_k^{rfsp} have completely known E-distributions (including known variances).

EVS 2010 is a statistical method for taking the data and using it to make statistical inference about the primary quantity of interest, $t_{1-\gamma}$. Specifically, EVS 2010 derives a one-sided upper confidence limit for $t_{1-\gamma}$ that we denote here by T*. Note that in conventional statistical terminology a confidence limit for an unknown percentile is called a tolerance limit.

RATIONALE FOR A SIMPLIFIED SCENARIO

The NOP trip setpoint problem is one specific scenario in which all the above characteristics of the EVS 2010 statistical model hold. In the NOP scenario, the ripples Q and flux shape φ are complex, high-dimensional quantities and T^0 is a complicated function of these quantities involving maximum and minimum operations. The observation errors ε_j^{soro} and ε_k^{rfsp} are accordingly also high-dimensional random variables, arising from computational (and possibly also science) errors in the physics codes SORO and RFSP. However, the validity of the mathematical and statistical theory of EVS 2010 does not rely on any of those complex features of the NOP problem. Indeed, the authors explicitly make the point that their method is applicable to many other situations. As long as the statistical model holds, the theory is correct and EVS 2010 should be applicable, no matter what kinds of quantities Q and φ are, what kind of function T^0 is, or what kinds of distributions the various random variables have (or how those distributions arise) in the specific application scenario.

It is therefore legitimate to test EVS 2010 in a scenario where these things are much simpler than in the NOP trip setpoint problem. In the Group A benchmarking scenario, the major simplification will be that Q and φ are simple scalar quantities. The T^0 function will also be very simple, and in particular will not involve extremal operations.

Working in a simplified scenario is not just a legitimate device but actually has some important advantages.

1. We can easily create test cases in which the true A-distribution of T is known, and in particular the true value of $t_{1-\gamma}$ is known.
2. We can simulate and analyse very large numbers of sets of data, so that the accuracy of the claimed tolerance limit property can be evaluated.
3. We can repeat these analyses using many variations on the problem, in order to explore the performance of EVS 2010 under a range of situations.
4. The simplicity also creates clarity. In a simple scenario it is easier to understand the behaviour of the method and to gain insight into the causes of good or bad performance.

Of course, working in a simplified scenario also has the major disadvantage that it is not the NOP problem. The primary objective of this project is to evaluate EVS 2010 for use in computing NOP trip setpoints. This raises an important question – what, if anything, can we learn about the performance of EVS 2010 in the NOP problem from exploring its performance in a simplified problem?

We cannot know whether in real applications the EVS 2010 method will perform in the same way against the evaluation criteria as it does in the benchmarking tests. In fact, some things will surely be different. But we do not have the opportunity to evaluate it in the real applications. In any real application we will not know the true values that we seek to place limits on, nor will we have the luxury of many repetitions.[1] In any real application we have just one set of data, which we analyse to give one result (one T* value) and we will certainly not know the true value $t_{1-\gamma}$ that we seek to bound.

This is a familiar situation. Every practical application of a statistical analysis faces the same issues. Assumptions are made that we know in practice will surely be false. Hopefully, reality will be close to the assumptions, but we cannot know whether it is. Approximations are made in any complex problem because we cannot derive an exact analysis. Hopefully, the approximations will yield answers that are close to those that a hypothetical exact analysis would produce, but we cannot know whether they are. True values of the things we are trying to make inference about are unknown, otherwise there would be no need to use the statistical method at all. We use approximations and make assumptions *precisely because* we cannot do anything better. The only guidance we have as to how a method will work in real applications is made up of

- theoretical properties that we know won't hold in reality, and
- some tests in simplified problems where we can see how well the method performs.

The guidance offered by simplified benchmarking tests is imperfect and fallible, but it is still worth doing. We know that performance in real applications will be different but we do not know in what respects and in what directions it will differ.

In general, our best guess of performance in reality is what we observe in the simplified problems.

The only exception to this rule is if we have (a) specific knowledge about how the simplified scenario differs from reality *and* (b) understanding of how those differences translate into specific performance differences. If,

[1] Even if the NOP trip setpoint calculation may be repeated on future occasions, those occasions will be separated by many months or years and the underlying reference value $t_{1-\gamma}$ will have changed. Indeed, it is the evolving conditions in the reactor that will prompt fresh computations. So we can consider each case as a fresh problem.

for instance, we observe undesirable performance in the tests then we must expect similar undesirable performance in serious applications. We can only argue otherwise if we have specific reasons based on knowledge of the method and the test scenarios. Conversely, if we observe good behaviour in the tests then we should expect similar good behaviour in reality. A critic who wishes to claim otherwise will need specific reasoning based on understanding of the method and the test scenarios.

THE BASE CASE

As mentioned above, in the Group A benchmarking tests Q and ϕ are simple scalar quantities. The T^0 function has the following form.

$$T^0(Q, \varphi) = 2 + aQ + b\varphi + cQ^2\varphi + d\varphi^2$$

where a, b, c and d are parameters that can be varied in the different test suites. Their values in the base test are

$$a = 0.1, \quad b = -0.2, \quad c = -0.03, \quad d = -0.4.$$

Although there are no extremal operations in this T^0 function, nonlinearity is introduced so as to create skewness in the A-distribution of T, even when the A-distributions of Q, φ and ϑ may be symmetric. The negative values for the parameters c and d lead to negative skewness, which is a feature that we expect to see in the NOP problem.[2]

Q is a random variable with the standard normal distribution, i.e. the normal distribution with zero mean and variance 1. This will be fixed in all the Group A tests. The distribution of φ, however, will vary between the different tests. In the base case, φ has 20 possible values, all having equal probabilities or weights. The possible values are evenly spread over the range [0, 1]; thus, the 20 values are 0.025, 0.075, 0.125, …, 0.925, 0.975.

To complete the specification of A-uncertainties, the distribution of ϑ is normal with zero mean and variance 0.0025. This will also be fixed in all the tests. The E-uncertainties are specified through the distributions of the other two error terms ε_j^{soro} and ε_k^{rfsp}. These are again normal distributions with zero means. The variances $var(\varepsilon_j^{soro})$ and $var(\varepsilon_k^{rfsp})$ are parameters that will be varied in the Group A tests. Their base case values are

$$var(\varepsilon_j^{soro}) = 0.04, \quad var(\varepsilon_k^{rfsp}) = 0.001.$$

The magnitudes of these variances were chosen with the following considerations. The standard deviation 0.05 for ϑ is relatively small compared with the variability in T that is introduced through T^0 and the uncertainties in Q and φ. This should reflect the fact that this error will be relatively small in the NOP problem. The base case standard deviation 0.2 for E-uncertainty in the sampled values of Q is also fairly small but non-negligible compared with the standard deviation of 1 in the underlying A-distribution of Q. This will introduce appreciable uncertainty about the A-distribution of Q even when the sample size N is large. The base case standard deviation 0.0316 for E-uncertainty in the estimated values of the φ_ks compares with the difference of 0.05 between neighbouring true values. It will lead to some clumping or even inversion in the order of the estimated values, whilst still being small compared with the overall range of [0, 1]. Again this reflects the expectation that RFSP code errors in the NOP problem will be relatively small.

[2] The constant 2 in T^0 produces values of $t_{1-\gamma}$ that are in most tests (and in particular in the base test) a little above 1. Although this is rather a trivial point, it does mean that the Group A tests are dealing with numbers that are in the familiar range of the NOP problem.

Another important feature of the base case is that where the statistical model assumes that parameters are known then this assumption is correct. For instance, the model assumes that the distributions of ϑ, ε_j^{soro} and ε_k^{rfsp} are known, including their variances. In the base case, the assumed values of these variances are equal to their true values. Similarly, the model assumes that the number and weights of the possible values for φ are known, and in the base case the assumed number and weights are the correct values.

TEST SUITES

Six test suites were specified, comprising 21 individual test cases. Together with the base case, there were 22 Group A benchmark tests in all.

SUITE 1. SHAPE OF THE TSP DISTRIBUTION

In this suite we vary the a, b, c and d constants to get A-distributions of T with greater skewness. This will test whether EVS 2010 retains good performance under a range of shapes. The rationale for this test suite is the concern expressed in my Review about the estimation of a key quantity $r_{1-\gamma}$ in the EVS 2010 methodology. This quantity measures the degree of skewness in the A-distribution of T. By varying the degree of skewness we can see whether this part of the EVS 2010 method performs well over a range of degrees of skewness.

Two variants of the baseline settings were considered.

1.1. a = 0.1, b = –0.1, c = –0.07, d = –0.5. The value of the standardised skewness measure k_3 (equal to the third central moment divided by the cube of the standard deviation) for this case is –0.8,[3] compared to –0.4 for the base case. However, the value of $r_{1-\gamma}$ is very similar at –1.72, and is not far from that of a normal distribution. The true value of $t_{1-\gamma}$ is also similar to the base test value at 1.37.

1.2. a = 0.1, b = 0, c = –0.03, d = –1. Although the skewness measure in this case is only k_3 = –0.6, the A-distribution of T is more non-normal and has a value of $r_{1-\gamma}$ = –1.83, close to that of the Gumbel distribution. The true value of $t_{1-\gamma}$ is 1.09.

SUITE 2. RELATIVE IMPORTANCE OF Q

As in Suite 1, we vary a,b,c and d, but now with the intention of increasing or decreasing the extent to which uncertainty about T is driven by random variation of Q as opposed to φ or ϑ. This will test how performance of EVS 2010 holds up when using data to learn about the Q distribution is more or less important. The rationale for this test suite is a purely statistical one. We have quite different forms of evidence with which to learn about the uncertainties in the A-distributions of Q and φ. In particular, whereas the amount of evidence about φ is fixed the amount of evidence about Q increases with the sample size, N. The tests in this suite will allow us to see whether EVS 2010 handles one form of uncertainty better than the other.

Again we have two variants on the baseline settings.

2.1. a = 0, b = –0.3, c = 0, d = –0.3. In this case the influence of Q becomes zero. This case was found to be quite revealing in the analysis of the test outcomes.

2.2. a = 0.3, b = –0.2, c = –0.05, d = –0.2. The influence of Q is increased here through the larger (absolute) values of a and c.

[3] Note that all `true values' quoted in this section are based on relatively small samples of 100,000 T values made while devising the test suites, and so are to some extent approximate (particularly the k_3 values).

Note that in both Suites 1 and 2, like in the base case, the assumed parameter values equal their true values, because the function T^0 is always known.

SUITE 3. DEGREE OF EPISTEMIC NOISE

In this suite we vary the E-distributions of ε^{soro} and ε^{rfsp} to get larger or smaller variances. The assumed variances will in every case match the true values, so that there is no specification error. This will test how well EVS 2010 performs when the data are more or less noisy. However, this suite has another important rationale. One of my major concerns in my Review was the existence of behaviour that I called paradoxical. A number of instances were reported to me in which as one or more variances of E-uncertainties were increased (or decreased) the NOP trip setpoint computed by EVS 2010 also increased (or decreased). Intuitively, with more (or less) noise in the observations the statistical estimation would become more (or less) cautious. If one or more E-uncertainty variances increase, therefore, we should expect to see a more cautious, i.e. lower, computed trip setpoint. Movements in the opposite direction I called paradoxical. At the time of completing my Review, it was not clear just when or why this behaviour arose. It was described as paradoxical to highlight the importance of gaining understanding into this phenomenon, in particular to determine whether it constituted reasonable or unreasonable behaviour.

My Review found that it was not clear whether the phenomenon arose only when changing the assumed values of variances, whilst keeping the data unchanged. Then the behaviour might be a reasonable result of using an assumed variance that does not agree with the true value underlying the data. But it might also arise when both true and assumed variances were changed, in which case the behaviour would be unreasonable. Test Suite 3 examines this latter situation.

This suite comprises five variants on the baseline settings of $\mathrm{var}(\varepsilon^{soro}) = 0.04$, $\mathrm{var}(\varepsilon^{rfsp}) = 0.001$.

3.1. $\mathrm{var}(\varepsilon^{soro}) = 0$, $\mathrm{var}(\varepsilon^{rfsp}) = 0.001$.
3.2. $\mathrm{var}(\varepsilon^{soro}) = 0$, $\mathrm{var}(\varepsilon^{rfsp}) = 0$.
3.3. $\mathrm{var}(\varepsilon^{soro}) = 0.04$, $\mathrm{var}(\varepsilon^{rfsp}) = 0.01$.
3.4. $\mathrm{var}(\varepsilon^{soro}) = 0.4$, $\mathrm{var}(\varepsilon^{rfsp}) = 0.001$.
3.5. $\mathrm{var}(\varepsilon^{soro}) = 0.4$, $\mathrm{var}(\varepsilon^{rfsp}) = 0.01$.

SUITE 4. MIS-SPECIFIED EPISTEMIC ERROR

This suite is the same as Suite 3, except that the true E-distributions stay fixed and only the assumed distributions change. The rationale for this is again two-fold. First it is interesting to see the extent to which the EVS 2010 tolerance interval performance is affected by mis-specification. In real applications, assuming known distributions, and in particular known variances, is quite a strong assumption. We must expect mis-specification in practice. By observing which kinds of mis-specification most affect performance we can be guided as to which values should be specified with the greatest care.

The second rationale is the issue of paradoxes mentioned in the discussion of Suite 3. In this suite, the assumed error variances are changed but the true values remain fixed.

This suite comprises six variants. Variants 4.1 to 4.5 are the same five variants as Suite 3, except that now the true error variances retain their baseline settings of $\mathrm{var}(\varepsilon^{soro}) = 0.04$, $\mathrm{var}(\varepsilon^{rfsp}) = 0.01$, and it is only the assumed variances that change. Variant 4.6 applies a smaller mis-specification with assumed variances for ε^{soro} and ε^{rfsp} of 0.0576 and 0.00144 respectively.

SUITE 5. FLUX SHAPE DISTRIBUTION

This suite is similar to Suite 1, in the sense that it explores the effect of varying the true A-distribution of T, but now instead of varying the T^0 function we vary the flux shape distribution. The true and assumed distributions of φ are the same (and discrete), but we vary the weights of the discrete flux shape set. By this means we can obtain more extreme distributions for T, involving higher degrees of skewness, bimodality, etc. So this is a more stringent test suite than Suite 1.

The suite comprises two variant distributions of φ. In each case the distribution is discrete with possible values 0.025, 0.075, …, 0.975, and these same values are the φ_k values used for generating the Φ_ks. Notice that the formulation of T^0 is such that higher values of φ are associated with lower T. Whereas the baseline case assigns equal probabilities (0.05, 0.05, …, 0.05) to these values, the variants assign them different sets of probabilities. These are both the true probabilities and the assumed weights (hence there is no mis-specification).

5.1. (0.0125, 0.0375, 0.075, 0.1, 0.1125, 0.1125, 0.1, 0.075, 0.0625, 0.05, 0.0375, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025). This is a quite strongly skewed distribution that will produce more skewness in T.

5.2. (0.03, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.1, 0.12, 0.11, 0.09, 0.07, 0.05, 0.03, 0.01, 0, 0, 0, 0.03, 0.03). The bulk of this distribution has some negative skewness which will tend to counteract the skewness introduced by the nonlinearity of the T^0 function. However, it has a lump of probability at high φ values that will produce a lump at low values in the T distribution that is just big enough to affect $t_{1-\gamma}$.

SUITE 6. MIS-SPECIFIED FLUX SHAPE DISTRIBUTION

In this suite we hold the assumed flux shape distribution fixed but vary the true A-distribution of φ. The assumed distribution is therefore based on Φ_k estimates being generated from the 20 base case φ_k values 0.025, 0.075, …, 0.975, and with equal weights, but the true distribution is different. The rationale for this test suite is the concern expressed in my Review about the assumption in EVS 2010 that the A-distribution is discrete, with estimates Φ_k available for all the possible values φ_k, and with known probabilities/weights. There are three parts to this assumption and it may therefore fail in three possible ways.

A. The true distribution may be continuous. In my Review, I expressed the opinion that this was likely to be the case, my understanding being that the assumed discrete values are not the only possible values but a relatively dense set of representative values that is intended to cover the range of possibilities. In this suite, the underlying true distribution is allowed to be continuous. However, when the discrete set for which we have estimates, together with their assumed weights, is indeed a good representation of the true distribution (as it is in test 6.1) we may expect to observe good performance of EVS 2010.

B. The discrete set of estimates may not cover the range of possible flux shapes adequately. The true distribution, whether discrete or continuous, has some parts of its range for which no representative values are in the observation set. Test 6.2 allows the true distribution to have more extreme φ values than any in the set $\{\varphi_k\}$.

C. The assumed weights may be wrong. The true distribution may be discrete, with the same set of possible values but different weights, or it may be continuous such that the assumed distribution does not approximate to the true distribution's shape. Both of these cases are explored in this test suite, in tests 6.3 and 6.4.

Because this suite involves mis-specification, we do not expect EVS 2010 always to perform well. The objective is to investigate which kinds of mis-specification have the largest impacts on its behaviour.

This suite comprises four variants, in which the true distributions of φ are as follows.

6.1. The continuous uniform distribution over [0, 1]. In this case, the assumed distribution is a simple discretisation of the true distribution. The true distribution has a slightly larger variance and allows slightly larger values of φ (between 0.975 and 1).

6.2. The discrete uniform distribution with probabilities 0.04 on each of the 25 values 0.025, 0.075, …, 0.975, 1.025, …, 1.225. This is the only case in which φ can take values outside [0, 1], and represents the situation where the assumed distribution excludes some extreme flux shapes.

6.3. The continuous beta distribution Be(1, 1.5). This distribution is not far from uniform but is somewhat skewed. It gives more weight to lower φ values than the assumed distribution.

6.4. The discrete distribution with probabilities (0.1, 0.1, 0.05, 0.05, 0.05, 0.05, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.05, 0.05, 0.05, 0.05, 0.1, 0.1) on the same φ values 0.025, 0.075, …, 0.975 as the assumed distribution. This case corresponds to a mis-specification in which the true distribution is appreciably more dispersed than the assumed one.

COMPUTATION

The specification of the Group A benchmarking exercise was completed by details of the computations required.

First, both EVS 2010 and the Bayesian comparator should be used to compute trip setpoint values T*, for a large number of simulated datasets, and these T* values should be compared with the true A-distribution of T, and in particular the true $t_{1-\gamma}$ value, in each test case using four output measures. The Bayesian comparator and the output measures are described later in this report.

The true A-distribution of T for each test was computed by simulation. Random values were sampled repeatedly from the true A-distributions of Q, φ and ϑ and the corresponding value of $T = T^0(Q, \varphi) + \vartheta$ evaluated. The resulting large sample of T values is a simulation of its A-distribution. The number of simulated values was not specified, but was required to be large enough to evaluate $t_{1-\gamma}$ to a suitable degree of accuracy. The AMEC NSS team that were implementing the EVS 2010 method used 10,000,000 simulations, whereas the CNSC team that were implementing the Bayesian comparator used 500,000. They produced values for $t_{1-\gamma}$ that agreed to at least two decimal places for each of the 22 tests.

In each test, random sets of data were generated, and a T* value computed for each dataset using both EVS 2010 and the Bayesian comparator. Each dataset comprised N simulated Sj values (by drawing N random values Q_j from the true A-distribution of Q and then adding random noise, with the appropriate assumed variance, ε_j^{soro} to each), and 20 simulated Φ_k values (by adding random noise ε_k^{rfsp}, with the appropriate assumed variance, to each of the assumed φ_k values). Data were generated with three values of the ripples sample size, N = 20, N = 100 and N = 500. For each sample size, a large number of random datasets were generated. The AMEC NSS team used 10,000, while the CNSC team used 2,000.

The Group A benchmarking proposal document also made some suggestions for reusing simulations in order to reduce noise for comparisons between methods and between tests. These were followed by the CNSC team but not by AMEC NSS. Nevertheless, the numbers of simulated datasets (particularly the larger number used by AMEC NSS) were sufficient to make unambiguous comparisons where required – see the discussion of simulation accuracy for the four output measures later in this report.

THE BAYESIAN COMPARATOR

OUTLINE

This section describes in some detail the Bayesian method that I developed to act as a comparator for EVS 2010 in the Group A benchmarking tests.

The Bayesian analysis is used to quantify the epistemic uncertainties (E-uncertainties) regarding unknown features of the A-distributions which determine the A-distribution of T, and hence $t_{1-\gamma}$. There are two A-distributions that are not fully known.

1. The A-distribution of the ripples Q is treated as completely unknown. The information we have about this distribution comes from a sample of values S_j that are assumed to be equal to a sample of values Q_j randomly sampled from the A-distribution of Q but subject to observation errors ε_j^{soro} for j = 1, 2, …, N. The distribution of ε^{soro} is assumed known.
2. The A-distribution of the flux shape φ is treated as partly unknown. It is assumed to be discrete, with possible values φ_k and probabilities (or weights) ω_k for k = 1, 2, …, K. The probabilities are assumed to be known but the possible values are unknown. The evidence we have is a set of values Φ_k that are assumed to equal the underlying possible values but subject to observation errors ε_k^{rfsp}. The distribution of ε^{rfsp} is assumed known.

The Bayesian comparator derives posterior distributions for the A-distribution of Q and the set of φ_k values. It then computes T* by a simulation method involving two nested loops.

- In the outer loop, a random draw is made from each posterior distribution. That is, a random distribution for Q is sampled from its posterior distribution and a random set of φ_k values are sampled from their posterior distribution.
- In the inner loop, using the sampled values from the outer loop, the A-distributions of Q, φ and ϑ are now known and can be simulated from. An identical simulation process to that used to compute the true setpoint $t_{1-\gamma}$ is applied; denote the computed value by t*.
- Many iterations of the outer loop are conducted, sampling fresh draws from the posterior distribution each time, thereby producing many sampled values t*. The result T* is the lower 5% sample value, i.e. that which has 5% of the sampled t* values below or equal to it and 95% above.

It remains to describe how the posterior distributions are obtained and sampled from. Two quite standard Bayesian models and analyses are used for this. The next subsection provides the technical details. A final subsection discusses the strengths and weaknesses of the Bayesian comparator in the benchmarking context.

TECHNICAL DETAILS

The two simple Bayesian models used in the Bayesian comparator are first described in general terms before considering how they are applied to build the Bayesian comparator.

THE EXCHANGEABLE MEANS MODEL

Suppose that we have observations $x_i = \mu_i + \varepsilon_i$, for i = 1, 2, …, n, where the means μ_i are unknown parameters, the errors ε_i are independent N(0, t) and t is known. Now suppose that the μ_is are themselves drawn from a population so that we can write $\mu_i = \xi + \eta_i$, where ξ is the population mean, the η_is are independent N(0, u) and u is known. Finally ξ is given a vague prior equivalent to assuming it has infinite prior variance.

This model says that the unknown parameters μ_i are related (and, technically, they are exchangeable). They are modelled as being drawn randomly from a common population. Initially we have no idea what values they might take because we have no idea what the mean of that population might be. But as soon as we observe some data we begin to learn about ξ and we expect subsequent data to be similar (within the bounds set by the variances t and u).

The analysis of this model is completely standard. The posterior distribution can be characterised as follows.

- The posterior distribution of ξ is N(x¯, (t + u)/n), where x¯ is the sample mean of the x_is.
- Given ξ, the posterior distribution of μ_i is N((tξ + ux_i)/(t + u), tu/(t + u)).

We are interested in the posterior distribution of the μ_is. The parameter ξ is only a convenience introduced to simplify the analysis. We can obtain the marginal distribution of the μ_is by integrating ξ out from the joint posterior distribution. However, if we simply want to draw random variables from this joint distribution (as for instance in the simulation procedure to derive T*), we can just use the two-stage analysis above. First draw a sample value of ξ from its distribution and then sample each of the μ_is from its conditional posterior distribution using the sampled value of ξ.

The above analysis assumes that both t and u are known. It is usual to allow at least u to be unknown (and in particular this is what we do in the benchmarking problems). It is possible to extend the Bayesian analysis formally to allow u to be treated as another uncertain parameter and to derive its posterior distribution. However, the Bayesian comparator adopts a simplified approach. The sample variance of the x_is is an unbiased estimator of t + u, so if we denote this sample variance by v then we estimate u by v – t. The simplified method just substitutes v – t for u in the above formulae.

SHRINKAGE

The exchangeable means model is the simplest one that exhibits the phenomenon known as shrinkage. If we didn't have the exchangeable model structure linking the μ_is to a common ξ, we would simply estimate each μ_i to be equal to the observed x_i. But in the exchangeable means analysis the posterior estimates (i.e. posterior means) of the μ_is are shrunk towards the estimate of ξ (which is their mean x¯). Specifically, if the range of the x_i values is R then the range of the posterior means is uR/(t + u), i.e. smaller by the factor u/(t + u). The amount of shrinkage depends on the relative magnitudes of the variances t and u. If the observation error variance is large relative to the variability in the μ_is, then the shrinkage is strong.

There is a rational explanation for shrinkage. Because the x_is are equal to the μ_is plus observation error, each has a variance t + u, whereas the μ_is have variance u. So if we used the observed x_i values as estimates (as in the non-exchangeable model) they would on average have too large a spread. So it makes sense to reduce that spread. Shrinkage estimators arise naturally in this kind of Bayesian analysis, but they have also been advocated in frequentist statistics, precisely because of this argument.

But is u/(t+u) the right amount of shrinkage? Remember that if we multiply a random variable by a then we multiply its variance by a^2. So since the variance of the x_is is t + u the variance of the shrunk estimates is

$$[u/(t + u)]^2 (t + u) = u^2/(t + u) .$$

The variance of the μ_is should be u, and this is less than u, so it seems that the Bayesian estimates shrink too much. The resolution of this conundrum comes from noting that the Bayesian analysis does not say that the μ_is are equal to those shrunk estimates, only that these are their posterior expected values. The posterior distribution says that they are distributed around those shrunk estimates with variances tu/(t + u). So when we sample from the posterior distribution we get sample values of the μ_is whose variance is increased by this amount. And this is exactly the right amount, because if we add $tu/(t + u)$ to $u^2/(t + u)$ we get just u.

THE BAYESIAN BOOTSTRAP

Suppose that we observe a sample of data $q_1, q_2, \ldots, q_n$ drawn from a distribution G that is unknown. A standard Bayesian model for this problem assigns a Dirichlet process prior distribution to G. This distribution is characterised by a prior estimate G_0 and a prior weight d_0. Then after observing the q_is the posterior distribution of G is again a Dirichlet process with posterior weight $d_n = d_0 + n$. The posterior estimate G_n has the form $(d_0G_0 + nG_n)/d_n$, where G_n is the empirical distribution from the data. That is, G_n is a discrete distribution assigning probability (1/n) to each of the observed data values $q_1, q_2, \ldots, q_n$. This is another completely standard Bayesian analysis.

For the Bayesian comparator we use the weak prior information case where $d_0 = 0$ (and then G_0 is irrelevant). The posterior distribution is now a Dirichlet process with estimate equal to the empirical distribution G_n and weight n.

The Dirichlet process is necessarily a complex mathematical construct, but it is simple to draw a random value from it. Remember that this is a posterior distribution for the unknown data distribution G, so a random value drawn from this distribution must itself be a distribution, g. Random draws from a Dirichlet process with a discrete estimate distribution are themselves discrete. Specifically, g will be a discrete distribution with possible values $q_1, q_2, \ldots, q_n$ but with randomly drawn probabilities $p_1, p_2, \ldots, p_n$. These probabilities themselves have a Dirichlet distribution (not a Dirichlet process distribution) D(1,1,…,1), and there is a very neat and simple algorithm to draw a random set of probabilities from this specific Dirichlet distribution.

The algorithm is called the stick-breaking algorithm because we can think of it in terms of breaking a stick at n – 1 random places. We draw n – 1 random numbers between 0 and 1, arrange them in increasing order, add 0 at the beginning and 1 at the end, and then the probabilities p_1 to p_n are the differences between successive values in this increasing sequence.

THE BOOTSTRAP AND THE BAYESIAN BOOTSTRAP

This is called the Bayesian bootstrap because it is similar to the more familiar (frequentist) bootstrap. The idea of the bootstrap is to represent uncertainty about the underlying distribution G by drawing random samples of size n from the original sample $q_1, q_2, \ldots, q_n$. This sampling is done with replacement, meaning that in a given bootstrap sample we can draw the same number twice, three times or even more. The bootstrap sample can be seen as an empirical distribution, which is now again a discrete distribution over possible values $q_1, q_2, \ldots, q_n$ but with probabilities n_i/n, where n_i is the number of times that the i-th item q_i is drawn for the bootstrap sample. (In contrast, the empirical distribution of the original sample gives every q_i the same probability (1/n).)

As we have seen, a sample from the Dirichlet process posterior is also a discrete distribution over possible values $q_1, q_2, \ldots, q_n$ but with probabilities sampled by the stick-breaking algorithm. In contrast to the regular bootstrap, whose probabilities can only take values 0, (1/n), (2/n), …, (n-1)/n, 1, the probabilities in the Bayesian bootstrap are not constrained and can take any values in the range 0 to 1. Analyses using the Bayesian bootstrap thereby generally produce smoother results than the ordinary bootstrap.

FLUX SHAPE UNCERTAINTY

We now describe the application of these two models, the exchangeable means and the Bayesian bootstrap, in the Bayesian comparator for the benchmarking exercise. First, consider the uncertainty about the distribution of flux shape. As explained above, the uncertainty here lies only in the values of the φ_ks, which are observed with error as the Φ_ks. Here we use the exchangeable means model. It is a bit of a stretch in this application because we don't really consider the φ_ks to be randomly sampled from some distribution, but they are expected to be of similar magnitude. And the exchangeable means model produces shrinkage so that the samples of φ_k values used in the three-stage Bayesian simulation should have the right amount of spread.

Formally, we identify the components of the model with those of the flux shape application as follows:

- The observations x_i are the Φ_ks, and the sample size n becomes K. (K = 20 in the Group A tests.)
- The true values μ_i are the φ_ks.
- The observation error variance t is the variance of ε^{rfsp}.

RIPPLES UNCERTAINTY

The uncertainty about ripples has two components. Imagine first that there was no observation error, so that we actually observed the sample Q_j values. We apply the Bayesian bootstrap for uncertainty about the underlying distribution G. For this part of the modelling we identify the observations q_i with Q_j and the sample size n is N.

Next, we account for the fact that the ripples are observed with error as S_j, and we use the exchangeable means model (which now applies without any caveats). As for flux shape uncertainty, we can identify the components of the theoretical model with the ripple application as follows:

- The observations x_i are the S_js and the sample size n becomes N.
- The true values μ_i are the Q_js.
- The observation error variance t is the variance of ε^{soro}.

So this Bayesian analysis is in two parts, and the algorithm for sampling a distribution g accordingly has two steps:

1. Sample the Q_j values using the exchangeable means model.
2. These become the Q_j values in the Bayesian bootstrap model and we sample the probabilities p_j using the stick-breaking algorithm.

VALUE OF THE BAYESIAN COMPARATOR IN THE BENCHMARKING EXERCISE

It is common to test new methods against some comparator, but there are different kinds of comparator. We might consider three types, which I call the gold, silver and lead standards.

- A gold standard is a method that performs as well as is theoretically possible. The new method cannot beat it. The interest in such a comparison is how close the new method comes to the performance of the gold standard.

- A silver standard is the best method that currently exists. The interest in such a comparison is whether the new method can beat the silver standard (and so become the new silver standard). If not, it may be better in some respects although worse in others.

- A lead standard is like a placebo in a drug comparison, or what is sometimes called a "straw man". It is a method so basic that its level of performance is minimal. The interest in such a comparison is to verify that the new method is at least better than the lead standard.

For solving the problem posed by the EVS 2010 statistical model there is no gold standard. We do not know what limits exist to performance. Nor is there a silver standard because there are no pre-existing methods which solve the problem, in the sense of producing a tolerance limit. The Bayesian comparator has been created simply to provide some sort of alternative method. It should not be considered a silver standard because it is not intended as a serious method in its own right. It has been developed quickly using rather crude Bayesian models, so should not be expected to compete with the EVS 2010 method that has benefited from a lengthy and quite intensive development process. Nevertheless, it should not be seen as a lead standard, either. Hopefully it will be somewhere between silver and lead, a method that could perhaps be viable but which should not be hard to beat.

Further insight into the comparator can be gained by looking at its limitations in relation to those of EVS 2010.

First, the Bayesian comparator was developed specifically for the simplified scenario of the Group A benchmark tests. The key point here is that it requires both Q and φ to be scalar random quantities. Although this is not a limitation in respect of the Group A benchmarking, it means that it could not be used for the Group B tests without first being developed to work with vector quantities.

Of more direct relevance to the Group A tests is the fact that we have already remarked in relation to EVS 2010 that every practical method will make simplifying assumptions and approximations. The Bayesian comparator is no exception. The approximations made by the two methods are quite different, however.

- EVS 2010 makes two principal approximations when solving the basic statistical model. First, several of its formulae rest on the central limit theorem. They will be exactly correct for infinitely large samples but only approximate in real applications with finite data. Second, it uses what the EVS 2010 Report calls a surrogate method to compute several quantities that are needed in the formulae. This is another kind of approximation which would only be exact in the absence of observation errors.
- The Bayesian comparator, like any Bayesian method, rests on the prior distribution. The method is exact but its performance will depend on the appropriateness of the prior distribution. Since the elements of the comparator's prior distribution come from the simple Bayesian models used, they are simplifications. We can therefore consider the resulting computations to be an approximation to what might be produced with a more carefully considered prior formulation.

One way to look at the implications of these approximations is to consider the statistical property known as consistency. A statistical estimator is consistent if, as the quantity of data approaches infinity, it converges in probability to the true value. In particular, a tolerance limit for $t_{1-\gamma}$ will be consistent if, as the sample size increases, T* becomes arbitrarily close to the true $t_{1-\gamma}$, for any data.

We first note that neither EVS 2010 nor the Bayesian comparator can be consistent as N tends to infinity. The reason is that N refers only to the data S_1, S_2, …, S_N which provide information about the A-distribution of Q. As N goes to infinity we can certainly hope to learn perfectly the distribution of Q, but the quantity of information about the φ_k values is fixed at one observation per φ_k. So as long as there is observation error in the flux shape estimates then no matter how large N is we can never know the A-distribution of φ perfectly, and consistency is not possible.

So now suppose that there is no error in flux shapes, i.e. $\mathrm{var}(\varepsilon_k^{rfsp}) = 0$. Will the methods now be consistent? I am not sure of the answer for EVS 2010, but I think not. The central limit theorem will become exact in the limit, but I think the surrogate method will not. The estimates produced by the surrogate method will, in my judgement, not be consistent. If my judgement is correct, the T* values produced by EVS 2010 will converge in the limit to some value for any data, but that value will not be $t_{1-\gamma}$.

In general, I would expect the Bayesian comparator not to be consistent, either. The Bayesian bootstrap is a consistent method, so if there were no observation error in the ripples then the Bayesian method would estimate the A-distribution of Q consistently and so consistently estimate $t_{1-\gamma}$. But I suspect the same could be said for EVS 2010 because I think the surrogate method would then become consistent. The simple exchangeable means analysis will in general not be consistent because it relies on the assumption of normal distributions in the prior distribution and the observation error. As it happens, these assumptions are true in the specific context of the Group A benchmarking tests, so I believe the Bayesian comparator will be consistent in any such test for which $\mathrm{var}(\varepsilon_k^{rfsp}) = 0$. If, however, the true A-distribution of Q is not normal then the T* values produced by the Bayesian comparator, like EVS 2010, will converge to a value that is not exactly equal to $t_{1-\gamma}$.

Note that the exchangeable means model is also used in the Bayesian comparator for flux shapes, and that in the Group A benchmark tests the underlying distribution of φ is far from normal. So this approximation in the Bayesian comparator can be expected to affect its performance in the tests adversely.

In summary, both methods are imperfect because they make approximations. But perfection (a gold standard) is not available. EVS 2010 is the only available method capable of producing tolerance limit solutions for the full complexity of the NOP trip setpoint problem, but the Bayesian comparator should provide an interesting contrast in the Group A benchmarking tests.

EVALUATION CRITERIA

OUTPUT MEASURES

The performance of EVS 2010 in the Group A benchmarking tests will be evaluated using four summary measures of the T* values produced in each test for each sample size and each method. To understand the output measures it will be helpful to review the basis of the tolerance limit approach to trip setpoints.

The ideal trip setpoint T is a random quantity, subject to A-uncertainty. It therefore has a probability distribution. In the Group A tests we know all the true A-distributions, and hence can derive (by Monte Carlo simulation) the A-distribution of T. In particular, we can compute the value $t_{1-\gamma}$ for which

$$\Pr(T < t_{1-\gamma}) = 1 - \gamma\ .$$

A method for calculating a computed trip setpoint from available data produces an estimate T*. Because the data are random (having E-uncertainty), T* is also a random quantity. Different methods will be characterised by T* having different probability distributions. A proper β/γ tolerance limit is a method with the property that

$$\Pr\ (T^* > t_{1-\gamma}) = 1 - \beta\ .$$

The Group A tests employ the usual specification that both β and γ are set to 0.95, or 95%.

Figure 1 illustrates this framework. The solid red curve is the A-distribution of T, while the dashed blue curve is the E-distribution of T*.[4] The black vertical line is drawn at the value $t_{1-\gamma}$, so that the area under the red curve to left of this point is 5%. The blue dashed curve represents a proper 95/95% tolerance limit, and so the area under this curve to the right of $t_{1-\gamma}$ is also 5%.

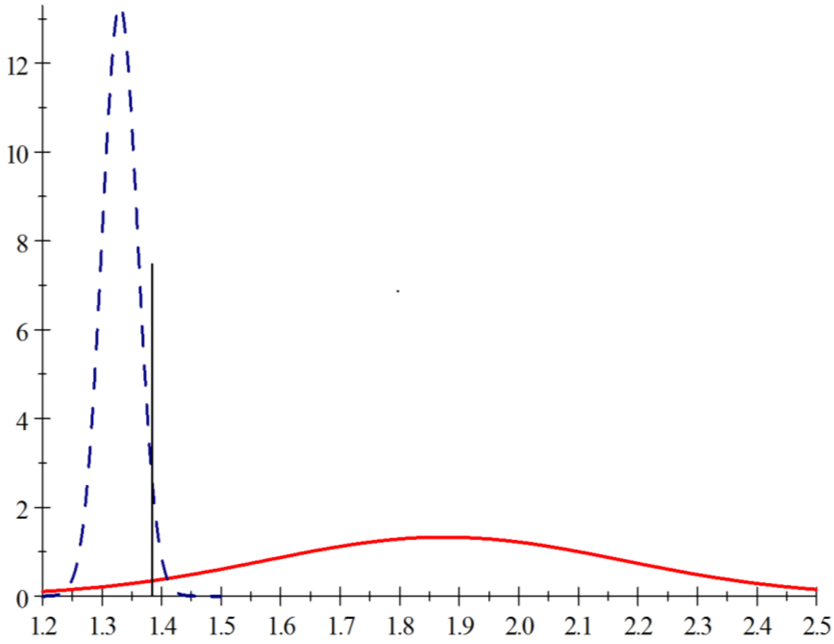

Figure 1. Example. A-distribution of T (red), E-distribution of T* (blue dashed) and $t_{1-\gamma}$

[4] I will use normal density functions for convenience in drawing figures, although these distributions will in practice not be normal (and in particular will typically be skewed).

Figure 2 shows a part of the same red A-distribution of T and three blue E-distributions for T* generated by three hypothetical estimation methods. The solid blue line is not a proper tolerance limit because the area under the curve to the right is more than 5%. This method would be undesirable because it generates trip setpoints that are generally too high. In particular, they are too often above $t_{1-\gamma}$.

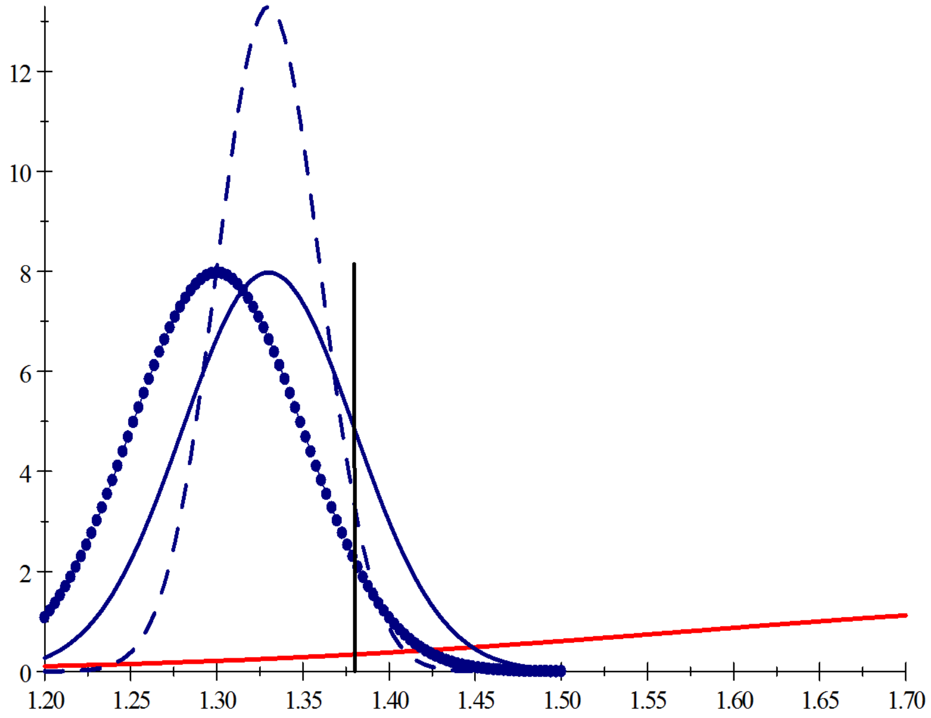

Figure 2. Example. A-distribution of T (red) and three E-distributions of T*

The dotted and dashed blue lines are both proper tolerance limits; both have area 5% to the right of $t_{1-\gamma}$. However, the method represented by the dashed line generates estimates T* with a lower standard deviation (SD) and would be preferred for two reasons. First, it has a higher mean T*, which means that on average it allows a higher operating power. For this reason it is preferred by the operator. Second, on the 5% of occasions when T* is higher than $t_{1-\gamma}$, it is on average less far above $t_{1-\gamma}$ than the method represented by the dotted line. For this reason it is preferred by the regulator.

Notice that the solid and dashed density functions both have the same mean, but the solid one is undesirable because it has a higher SD, leading to excessive probability of T* being above $t_{1-\gamma}$. For a proper tolerance limit to attain a higher mean T*, it must generally have a smaller SD.

The 4 measures used to benchmark the EVS method are as follows.

1. The *mean* of T*. All other things being equal, a higher mean is preferred.
2. The *SD* of T*. All other things being equal, a lower SD is preferred because it indicates that the method is making better use of the data, and because it tends to be associated with improvements in the other measures.
3. The *non-coverage*, which is the percentage area under the T* density curve to the right of $t_{1-\gamma}$. Ideally this should be 5%. In reality, all methods make approximations, so we cannot expect to achieve this ideal exactly. It is desirable for the non-coverage to be not greater than 5%.
4. The *mean deficit*, which is derived as follows. For every instance when T* is above $t_{1-\gamma}$, we compute the probability under the red curve to the left of T*. The mean deficit is the average of these percentages. It

will be 5% for any method for which T* never exceeds $t_{1-\gamma}$, and otherwise will be greater than 5%. All other things being equal, a smaller mean deficit is preferred.

EVALUATION CRITERIA

The four measures form the basis of the criteria that I will use to evaluate EVS 2010. The various Group A tests will be used for benchmarking EVS 2010 in two different senses. The first is to benchmark EVS 2010 alone as a formal statistical tolerance limit, while the second will contrast its performance with that of the Bayesian comparator. We could characterise these as *absolute* benchmarking and *relative* benchmarking.

There are five evaluation criteria for absolute benchmarking of EVS 2010.

1. The principal evaluation criterion for the EVS 2010 results alone is whether the EVS method is performing as a valid 95/95 tolerance limit. This is assessed by looking at the *Non-coverage* output in those tests (Suites 1, 2, 3 and 5) in which there is no deliberate mis-specification. Non-coverage should ideally be 5% in these tests, but we recognise that all practical methods are imperfect so we do not expect non-coverage to be exactly 5%. We would like it to be close to 5% and all other things being equal would prefer it to be less than 5% rather than greater.

2. Another useful measure is the *Mean deficit*. The primary criterion of non-coverage ensures that there is a suitably small probability that T* will exceed $t_{1-\gamma}$, while mean deficit is concerned with by how much it exceeds $t_{1-\gamma}$ on average when it does exceed it. The interpretation of mean deficit is somewhat convoluted because a tolerance limit involves two probabilities. A valid 95/95 tolerance limit will have only a 5% probability of producing a trip setpoint T* which is *excessive*, in the sense that in subsequent repeated use it would be too high to trip on more than 5% of future instances. A mean deficit value of 10% would mean that when T* is excessive in that sense, then in repeated use it would fail to trip on 10% of future instances. We would like mean deficit to be only a little above 5%.

3. An additional evaluation criterion for the EVS 2010 results is 'face validity', which is to assess whether as we vary some of the parameters of the tests the EVS 2010 results change in ways that accord with intuition. For this purpose we look particularly at the *Mean* and *SD* of the T* values. For instance, as the sample size increases we have more information. We should therefore expect that increasing sample size is associated with decreasing SD. In turn, decreasing SD should mean that a valid tolerance limit will have increasing mean.

4. Another important aspect of face validity is whether when epistemic error variances are varied we find behaviour that has been called *paradoxical* in my report on EVS 2010. Specifically, intuition suggests that if there is less epistemic uncertainty (decreased variances) then there should be less uncertainty about the true $t_{1-\gamma}$, and hence the computed tolerance limit should be higher. However, there was evidence that the EVS 2010 method could produce the opposite behaviour: I described this as paradoxical. AMEC NSS defended the behaviour, arguing that it was a natural and valid consequence of the use of extremal computations (maxima or minima) in the definition of T. Despite some discussion regarding this, the details of the argument remained obscure to me, and in particular I was not sure whether the paradoxical behaviour arose when the true epistemic variances were varied or only when the estimated variances were altered. One purpose of Suites 3 and 4 was to see whether this behaviour could be replicated in the simple abstract context of the Group A tests.

5. Finally, it will be interesting to see how robust or sensitive the EVS 2010 results are to the mis-specifications induced in Suites 4 and 6. We would not of course require the method to perform correctly

in the face of mis-specification; for instance non-coverage values quite far from 5% may be expected in some instances and will not in any way invalidate the method. However, any real application of any statistical method will make assumptions that do not hold precisely. A method which is robust to mis-specification will have performance that does not degrade much under realistic mis-specification.

Many of the evaluation criteria for relative benchmarking of EVS 2010 against the Bayesian comparator are directly related to the five criteria for absolute benchmarking. Thus, we can ask whether the Bayesian method succeeds in having non-coverage values close to 5% (or whether it generally has non-coverage closer or further from 5% than EVS 2010) – criterion 1. We can also ask whether it has comparably small mean deficit – criterion 2 – and whether it has face validity (both in terms of the way the mean and SD of T* behave as sample size increases and in terms of paradoxes) – criteria 3 and 4. And we can ask whether it is more or less robust to mis-specification of assumptions – criterion 5. However, we can also identify another important criterion for this comparison.

6. We wish to see whether EVS 2010 makes good use of the available information by comparing it with another method, the Bayesian comparator, that uses the same information. This is a judgement of relative efficiency. The principal basis for this comparison will be the *Mean* of T* values. A method that makes better use of the data should have a smaller SD, so this is also a measure of efficiency, but the important consequence of the smaller SD is that it allows a higher mean. A method which has smaller SD but the same mean is failing to make good use of the additional information.

RELATIONSHIP TO THE NOP TRIP SETPOINT PROBLEM

We have already remarked that the Group A benchmarking scenario is a simplified and abstract problem that is quite different from the NOP trip setpoint problem. But we also remarked that this kind of benchmarking is nevertheless valuable and may constitute the best guide we have to behaviour in the NOP problem, given that we cannot know true values or perform multiple simulations in the larger problem. In this section, I tentatively suggest what kinds of behaviour, as revealed by the evaluation criteria in the Group A tests, might indicate concerns for the use of EVS 2010 in the NOP problem.

Consider the individual evaluation criteria.

Non-coverage. We do not expect EVS 2010 to be a perfectly valid tolerance limit in practice, in the sense of having non-coverage always exactly 5%. Every statistical method makes assumptions, approximations and simplifications – generically, compromises – which will mean that in practice it does not have precisely the properties that the theory says it should have. This is why the first evaluation criterion does not ask for exactly 5% non-coverage but instead applies the weaker condition of non-coverage close to 5%. There is no hard numerical condition for what constitutes close enough. As a rough rule of thumb, I would suggest that non-coverage that is not within a factor 2 of the target, i.e. below 2.5% or above 10%, would be worthy of notice. A method whose non-coverage is outside these bounds in many of the tests is a cause for concern, although low non-coverage might be less worrying because it is erring on the safe side.

Mean deficit. A method that has excessive (e.g. greater than 10%) non-coverage may redeem itself by a mean deficit that is small, only a little above 5%. If we find mean deficit values of, say, 10% or more then this would certainly be cause for concern, even though it occurs in the abstract, simplified Group A context.

Face validity. The two face validity criteria are important. A method that does not satisfy a face validity criterion is behaving in a way which intuition says it should not. In a sense, finding such problems in a simple, abstract example would be particularly worrying unless an explanation could be found to suggest that the behaviour would be different in a real application. Or unless we could show that the original intuition was false. Without such mitigations, a method whose behaviour does not have face validity should **fail** the benchmarking tests because we cannot have confidence that it will not produce nonsensical or seriously poor results in real applications.

Robustness. The fifth evaluation criterion is not clear cut. The purpose of stressing a method by stipulating that it must make assumptions which are false is not to see whether problems arise, since problems are certainly to be expected. It is to see *where* the problems arise. If we know which kinds of mis-specification cause the method to perform badly then this gives some guidance about which assumptions will be most important in a real application. The guidance is imperfect because the real application will be different in many ways from the benchmarking tests, but it is the best guidance we can get.

Efficiency. The final evaluation criterion applies only to relative benchmarking. The Bayesian comparator has been developed quickly, whereas EVS 2010 is the fruit of some man-years of effort, and on this basis it would be a reasonable expectation that EVS 2010 will be found to be more efficient in the benchmarking tests. But we should remember that different methods will usually have different strengths and weaknesses. One method may be particularly efficient in the context of the group A benchmarking tests but lose that superiority in real applications such as the NOP trip setpoint problem. Nevertheless, in the absence of specific reasons to suppose that performance in the real application will differ in a particular way the benchmarking tests are the best guide we have.

Finally, any unexpected or erratic behaviour would be cause for concern because it would make performance in real applications unpredictable. Even if performance against the evaluation criteria is generally good in the benchmarking tests, unexpected or erratic behaviour casts doubt on whether that good performance will hold also in real applications.

ACCURACY OF RESULTS

The four output measures are estimated from the large set of simulated T* values (10,000 for EVS 2010 and 2,000 for the Bayesian comparator) that are output by the test programs. Thus, the mean and SD of T* are estimated by the mean and SD of the simulated T* values. The non-coverage is estimated by the proportion of simulated T* values that exceed the true $t_{1-\gamma}$. And the mean deficit is estimated by finding, for each simulated T* that exceeds $t_{1-\gamma}$, the proportion of simulated T values that it exceeds, and then averaging those proportions. The accuracy of these estimates is limited by the number of simulations performed.

IDEAL ACCURACY

First consider the accuracy of the reported results assuming that in each case T* has been computed with perfect accuracy.

The results from both EVS 2010 and the Bayesian comparator are based on many simulated samples of S and Φ values. The Bayesian results are derived from 2,000 simulations. This was the specified number of simulations according to the “First Benchmarking Round Proposal (Draft 4)”. That number was chosen so that if the methods are valid tolerance limits, so that the true non-coverage value is 5%, then the standard error of the reported non-coverage (being the proportion of non-coverage in 2000 simulations) would be 0.005, or half of one percent.[5] So the Bayesian non-coverage values are accurate to this level, and any reported non-

[5] This calculation assumes not only that T* is computed perfectly in each simulation but also that $t_{1-\gamma}$ has been computed perfectly. This is a reasonable assumption since $t_{1-\gamma}$ was computed using a very large number of simulated T values (10^7 in the case of the EVS results).

coverage of 6% or more can be taken to denote a true non-coverage that is above 5%. Similarly, any reported value below 4% is evidence that the true non-coverage is less than 5%.

The EVS 2010 results were in fact computed from a larger number of simulated samples, i.e. 10,000. This means that the EVS reported non-coverage figures have a standard error reduced to 0.002. Therefore, any EVS 2010 reported non-coverage of 5.5% or more denotes a true non-coverage of over 5% (and a reported 4.5% or less denotes a true value below 5%).

We can also examine the accuracy of the estimated mean T* values. The accuracy varies with the SD of the T* values in the same test. The standard error of the estimated mean is 0.01 times the SD for the EVS 2010 outputs and about 0.025 times the SD for the Bayesian comparator. For the SD values a rough calculation suffices to show that they are accurate to plus or minus one percent or better.

The accuracy of mean deficit values is harder to assess. However, when the non-coverage is 5% or more the mean deficit values are based on at least 500 instances for EVS results or 100 for Bayesian results. The standard deviation of deficit values was not reported, but we can take it to be no more than the difference between the estimate and 5%. So, for instances where non-coverage is 5% or more the standard error of reported mean deficit percentages is at most 0.25 (EVS 2010) or 0.5 (Bayesian comparator) when the reported mean deficit is 10%. When the reported mean deficit is 6% these standard errors fall to 0.05 (EVS 2010) or 0.1 (Bayesian comparator).

REAL ACCURACY

The above analysis is described as relating to 'ideal accuracy' because it assumes that T* is computed perfectly. In reality, it is not an exact calculation in either method.

The EVS 2010 method employs a so-called 'surrogate method' to estimate some of the quantities required for computing T*. The accuracy of these estimates is limited by the number of internal simulations employed for this estimation, which was 10,000.[6]

The Bayesian computation also uses simulation, in this case involving two nested loops. The accuracy is limited by the number of simulations in the outer and inner loops, which was 7000 in both.

The effect of inaccurate computation of T* is to inflate the SD values. This in turn will inflate the non-coverage values. So reported values of both SD and non-coverage will be to some extent biased upwards. However, the numbers of simulations used (10,000 in EVS 2010 and 7,000 in both loops in the Bayesian method) are relatively large and were chosen with a view to obtaining stable T* computations, so I believe this bias will be small and can be ignored. Real accuracy will in practice be almost the same as ideal accuracy.

[6] Note that the surrogate calculations are also averaged over the sample S values, and their accuracy is therefore also limited by the sample size – 20, 100 or 500. However, this limitation is intrinsic to the EVS 2010 method because it must work with whatever sample size is available.

ANALYSIS OF OUTPUTS

THE RESULTS

The Group A benchmarking exercise comprises a baseline test and 6 suites of variants on that baseline, comprising a total of 22 tests. Each test was conducted using both the EVS 2010 method and a Bayesian comparator. Each test was run using 3 different sample sizes – 20, 100 and 500. For each test, sample size and method, 4 output measures were reported.

I was able to examine the code used to implement both the EVS and Bayesian methods, and can verify that they applied all the Group A benchmarking tests correctly.

The outputs are set out in Table 1. This has 22 blocks (the 22 tests) of 3 rows each (the 3 sample sizes). The results are presented in 4 blocks (the 4 measures) of 2 columns each (the 2 methods). The true TSP value $t_{1-\gamma}$ is also given for each test.[7]

TABLE 1: GROUP A BENCHMARK TEST RESULTS

Test		Mean		SD		Non-coverage		Mean deficit	
(True $t_{1-\gamma}$)	N	EVS	Bayes	EVS	Bayes	EVS	Bayes	EVS	Bayes
Base	20	0.995	1.3063	0.139	0.0631	0	9.1	5	5.94
(1.3791)	100	1.204	1.3376	0.055	0.0243	0	4.2	5	5.47
	500	1.286	1.3487	0.026	0.0168	0	3.05	5	5.32
1.1	20	0.816	1.2301	0.208	0.1130	0	10.85	5	6
(1.3455)	100	1.082	1.2866	0.075	0.0365	0	4.1	5	5.5
	500	1.191	1.3079	0.035	0.0205	0	2.65	5	5.32
1.2	20	0.526	0.9707	0.194	0.0526	0	3.5	5	5.73
(1.0588)	100	0.829	0.9864	0.079	0.0403	0.04	2.7	5.6	5.65
	500	0.951	0.9893	0.042	0.0397	0.33	3.35	5.5	5.54
2.1	20	1.214	1.4086	0.070	0.0169	0	4.1	5	5.49
(1.4377)	100	1.355	1.4086	0.031	0.0169	0.11	4.15	5.6	5.49
	500	1.409	1.4086	0.017	0.0168	4.2	3.9	5.6	5.51
2.2	20	0.689	1.0502	0.285	0.2110	0.73	18.7	6.4	7.62
(1.2294)	100	1.001	1.1132	0.101	0.0818	0.58	7.2	5.7	5.79
	500	1.123	1.1733	0.042	0.0323	0.36	4	5.4	5.34
3.1	20	0.998	1.3089	0.138	0.0627	0	9.2	5	5.94
(1.3791)	100	1.206	1.3384	0.054	0.0237	0	3.6	5	5.49
	500	1.286	1.3489	0.026	0.0166	0.01	3.25	5.1	5.3
3.2	20	1.002	1.3143	0.137	0.0636	0	11.8	5	6.01
(1.3791)	100	1.205	1.3473	0.052	0.0213	0	5.75	5	5.37
	500	1.286	1.3627	0.022	0.0083	0	2.3	5	5.14
3.3	20	1.001	1.2593	0.152	0.0691	0.04	2.85	5.4	5.9
(1.3791)	100	1.206	1.2745	0.073	0.0549	0.39	1.75	5.8	5.88
	500	1.287	1.2776	0.049	0.0547	2.58	2.1	5.9	5.77

[7] The values given here are those obtained by CNSC; AMEC NSS obtained essentially the same values.

Test		Mean		SD		Non-coverage		Mean deficit	
(True $t_{1-\gamma}$)	N	EVS	Bayes	EVS	Bayes	EVS	Bayes	EVS	Bayes
3.4	20	0.981	1.2917	0.155	0.0768	0	11.2	5	6.23
(1.3791)	100	1.194	1.3332	0.060	0.0291	0	5.55	5	5.62
	500	1.278	1.3476	0.028	0.0174	0	3.25	5	5.33
3.5	20	0.987	1.2471	0.170	0.0773	0.06	3.3	5.5	5.9
(1.3791)	100	1.198	1.2710	0.078	0.0557	0.37	1.8	6.1	5.87
	500	1.281	1.2756	0.052	0.0545	2.21	1.8	5.8	5.77
4.1	20	0.992	1.3038	0.139	0.0647	0	8.85	5	5.9
(1.3791)	100	1.200	1.3349	0.055	0.0246	0	3.4	5	5.43
	500	1.282	1.3461	0.026	0.0166	0	2.15	5	5.29
4.2	20	0.989	1.3068	0.141	0.0668	0	11.75	5	6.02
(1.3791)	100	1.197	1.3409	0.055	0.0254	0	6	5	5.55
	500	1.279	1.3569	0.026	0.0161	0	8.55	5	5.34
4.3	20	1.033	1.2888	0.137	0.0535	0	1.1	5	5.49
(1.3791)	100	1.232	1.3067	0.053	0.0225	0.05	0	5.5	5
	500	1.309	1.3106	0.025	0.0176	0.17	0	5.3	5
4.4	20	1.044	1.3372	0.135	0.0559	0.02	23.3	6	6.33
(1.3791)	100	1.242	1.3645	0.052	0.0234	0.08	26.2	5.5	5.73
	500	1.318	1.3722	0.025	0.0174	0.44	35.7	5.4	5.58
4.5	20	1.084	1.3125	0.131	0.0466	0.08	5	5.9	5.57
(1.3791)	100	1.271	1.3274	0.051	0.0221	0.99	0.7	5.6	5.4
	500	1.344	1.3298	0.025	0.0179	8.01	0.25	5.6	5.21
4.6	20	0.999	1.3062	0.141	0.0617	0	8.75	5	5.87
(1.3791)	100	1.207	1.3363	0.054	0.0239	0	3.55	5	5.45
	500	1.288	1.3463	0.026	0.0168	0	2.35	5	5.3
5.1	20	1.177	1.3988	0.120	0.0446	0.06	8.6	5.9	5.61
(1.4543)	100	1.357	1.4217	0.047	0.0194	1.01	4.45	5.6	5.35
	500	1.425	1.4292	0.022	0.0139	7.44	3.45	5.5	5.23
5.2	20	1.219	1.4333	0.116	0.0486	0.04	13.35	5.4	5.62
(1.4803)	100	1.384	1.4451	0.045	0.0226	0.97	5.6	5.3	5.38
	500	1.448	1.4518	0.021	0.0164	6.08	4.05	5.3	5.22
6.1	20	0.995	1.3063	0.139	0.0631	0	9.8	5	5.91
(1.378)	100	1.204	1.3376	0.055	0.0243	0	4.4	5	5.49
	500	1.286	1.3487	0.026	0.0168	0	3.55	5	5.32
6.2	20	0.995	1.3063	0.139	0.0631	13.55	98.05	5	15.81
(1.1424)	100	1.204	1.3376	0.055	0.0243	86.59	100	6.4	17.76
	500	1.286	1.3487	0.026	0.0168	100	100	10.5	18.59
6.3	20	0.995	1.3063	0.139	0.0631	0	0.05	5	5.07
(1.464)	100	1.204	1.3376	0.055	0.0243	0	0	5	5
	500	1.286	1.3487	0.026	0.0168	0	0	5	5
6.4	20	0.995	1.3063	0.139	0.0631	0.01	38.5	5.4	6.97
(1.33)	100	1.204	1.3376	0.055	0.0243	0.47	63.35	5.6	6.22
	500	1.286	1.3487	0.026	0.0168	3.57	87.35	5.5	6.24

ANALYSIS OF RESULTS – EVS 2010

I will present my analysis of the results in two parts according to the two forms of benchmarking. The first is absolute benchmarking, assessing the performance of EVS 2010 in isolation against the first 5 evaluation criteria.

VALIDITY AS A TOLERANCE LIMIT

The first, and most fundamental, finding is that the EVS 2010 non-coverage values are below 5% in the great majority of tests. Indeed, in many cases they are so low that their estimates are zero.

We see that the non-coverage is well above 5% in test 6.2. However, this should not be considered as a failure because test 6.2 is one of the tests (Suites 4 and 6) in which the methods are deliberately applied with mis-specified information. If assumed values of quantities are seriously wrong then no method can be expected to perform as it should do when the assumed values are correct. Indeed, it is particularly noteworthy that EVS manages to keep the non-coverage below 5% in every other test in Suites 4 and 6, except for test 4.5 with N = 500.

Evaluation criterion 1 suggests that we would like to see non-coverage values in the range 2.5% to 10%. EVS 2010 generally produces values below 2.5%. In the four test suites in which there is no mis-specification, the non-coverage only exceeds 2.5% in 5 cases out of 36 (12 tests and 3 values of N). The suggestion is that EVS 2010 is generally tending to under-predict, a point to which I will return in later discussions. Only 2 of these 5 cases showed non-coverage values over 5%, and it is noteworthy that these are both in test suite 5. Indeed, for N = 500 the non-coverage is above 5% for both of the tests in Suite 5. There is no mis-specification in these tests. They are characterised by flux shape distributions that are uneven, strongly skewed or bimodal, raising the question of whether this is just chance or whether there is some particular sensitivity in EVS 2010 to the shape of the A-distribution of φ.

In general, we notice that the non-coverage increases steadily with the sample size N, but instances of non-coverage below 2.5% are by no means confined to the smaller N values; they are in the majority (7 out of 12) even when N = 500. On the other hand, increasing non-coverage raises questions about what happens when N becomes much larger. There is no reason for concern if non-coverage asymptotes at or close to 5%, but notice that Suite 5 already indicates that it will not necessarily stay below 5%. It is possible that with higher sample sizes the non-coverage may go above 5% as a general rule.

In summary, EVS 2010 does not achieve non-coverage that is generally close to 5%, the great majority of reported values (in tests without mis-specification) being outside the range of 2.5% to 10%. However, there are no instances of non-coverage exceeding 10%; all the values that are not close to 5% are below 2.5%, with many being reported as zero, and values below 2.5% err on the safe side and therefore cause less concern overall.

Evaluation on Criterion 1 (validity as a tolerance limit) – Some questions have been raised about specific aspects of its performance on this evaluation criterion, but generally EVS 2010 performs satisfactorily in respect of the tolerance limit property.

MEAN DEFICIT

Although non-coverage that deviates from 5% may strictly mean that a method is not a valid tolerance limit, we see that the mean deficit for EVS is always below 6% (except in test 6.2 and in a couple of other instances in which N is less than 500). For instance, if the non-coverage were to be 10% the method is not in this instance a valid 95/95 tolerance limit and could claim only to be a 90/95 limit. But if the mean deficit is only 6% then it is most probably a valid 95/94 tolerance limit or better.

Computations of this kind were actually reported by NSS in addition to the four measures that were required of them. For example, in test 4.5 with N = 500 the EVS 2010 non-coverage is 8%. One way of looking at this is that it has failed the criterion for a valid 95/95 tolerance limit and qualifies only as a 92/95 tolerance limit. However, the mean deficit is only 5.6% and NSS report that it can also be considered a 95/94.7 limit. So when a method fails to satisfy the 95/95 tolerance limit criterion the non-coverage perhaps exaggerates the extent to which it misses that target.

In general, for a given non-coverage value, the mean deficit will decrease with decreasing SD. So the concern expressed above that increasing sample size may lead to non-coverage exceeding 5% is tempered by the fact that increasing sample size will also lead to smaller SD.

It is not clear whether the result will be to contain mean deficit to only a little above 5% even if non-coverage does increase above 5% with increasing N, but such indications as we have from the Group A results is that it will.

Evaluation on Criterion 2 (mean deficit) – EVS 2010 has excellent performance on this evaluation criterion.

MEAN AND SD

As the sample size increases, the SD of T* decreases and the mean increases towards (but except in test 6.2 never exceeds) $t_{1-\gamma}$. EVS 2010 therefore satisfies this aspect of the face validity criterion.

Evaluation on Criterion 3 (basic face validity) – EVS 2010 has basic face validity in respect of the behaviour of the mean and SD of T* with increasing N, and so passes the test of this evaluation criterion.

PARADOXES

The second aspect of face validity in the evaluation criteria is the avoidance of paradoxes as epistemic error variances are varied. The epistemic error variances in the base case are var(ε^{soro}) = 0.04, var(ε^{rfsp}) = 0.001. In the five tests of Suite 3 these were varied as follows.

3.1. var(ε^{soro}) = 0, var(ε^{rfsp}) = 0.001.

3.2. var(ε^{soro}) = 0, var(ε^{rfsp}) = 0.

3.3. var(ε^{soro}) = 0.04, var(ε^{rfsp}) = 0.01.

3.4. var(ε^{soro}) = 0.4, var(ε^{rfsp}) = 0.001.

3.5. var(ε^{soro}) = 0.4, var(ε^{rfsp}) = 0.01.

Consider first the effect of varying ε^{soro}, as shown by the mean T* values in Table 1. Comparing the base case with test 3.1, we see that reducing just ε^{soro} increases the mean T*. Comparing the base case with test 3.4, or test 3.3 with test 3.5, we see that increasing just ε^{soro} reduces the mean T*. These results agree with intuition.

Now consider the effect of varying ε^{rfsp}. Comparing test 3.1 with test 3.2, we see that reducing just ε^{rfsp} increases the mean of T*, which is again in accord with intuition. However, comparing the base case with test 3.3, or test 3.4 with test 3.5, we find that increasing just ε^{rfsp} also increases the mean T*. This does not agree with intuition. Although the changes are small, the sample sizes are reasonably large, and the consistency of the change across these two between-test comparisons and across three sample sizes suggests that this is a real effect, not simply sampling noise. Table 2 confirms this finding statistically.

TABLE 2: STATISTICAL SIGNIFICANCE OF PARADOXES

Comparison	N	Mean diff	SE
Base v 3.3	20	0.006	0.0021
	100	0.002	0.0009
	500	0.001	0.0006
3.4 v 3.5	20	0.006	0.0023
	100	0.004	0.0010
	500	0.003	0.0006

For each of the two comparisons and three sample sizes, the mean difference is the difference in mean T* values (positive values being increases in mean, contrary to intuition). The column labelled SE is the standard error of this mean difference.[8] The means have been reported only to three decimal places, and if they had been reported to four places the difference might have been different by up to 1 in the third place. But even with the most favourable assumption of subtracting up to 0.001 from all the mean differences, they are all positive[9] and four of the six exceed two standard errors (indeed, two equal or exceed three standard errors). The statistical analysis is quite emphatic.

Suite 4 examines what happens when assumed epistemic error variances are mis-specified. Tests 4.1 to 4.5 are the same as tests 3.1 to 3.5 except that only the assumed variances are changed – the true underlying variances remain the same as in the base case. Intuition is not quite so clear in these tests as in Suite 3. On the one hand, if we believe there is less epistemic error then we should be more confident about estimating $t_{1-\gamma}$, resulting in a higher trip setpoint. However, if there really is more epistemic error than stated, this might cause the tolerance limit to move in the other direction.

If we adopt the view that decreasing assumed error variances should increase $t_{1-\gamma}$, then we see paradoxical behaviour in *every one* of the six between-test comparisons analogous to the Suite 3 comparisons discussed above. The directions in which the mean T* moves are consistently opposite to that prediction across all six comparisons and all three sample sizes. Furthermore, the changes are appreciably larger than the paradoxical changes found in Suite 3. An additional test, 4.6, increased both assumed epistemic error variances by more modest factors and again we see increased mean T* values compared with the base case. Nevertheless, it is not clear whether such behaviour should be regarded as paradoxical.

[8] The standard error on an individual mean is 0.01 times the SD. To get the SE of the mean difference, we square the individual standard errors, then square-root the sum of those squares.

[9] Even in the comparison of base versus 3.2 and N = 100, a difference of 0.001 when rounded to three decimal places might have been very small before rounding but must nevertheless have been strictly positive.

It should be noted that the paradoxical behaviour in Suite 3 (and perhaps in Suite 4) arises without the Group A tests having any extremal functions in the definition of T^0.

I regard the paradoxical behaviour in Suite 3, when var(ε^{rfsp}) is varied, as a strict violation of face validity. Unless reasons can be brought forward to justify this behaviour or to show it will not happen in more complex problems such as the NOP trip setpoint problem, this is a serious matter of concern. Any violation of face validity damages confidence in a method. There is no assurance that it will behave sensibly in other applications.

Evaluation on Criterion 4 (paradoxes) – the mean T* values reported for EVS 2010 behave counter to intuition when the variance of ε^{rfsp} is varied in Suite 3 (without mis-specification). Pending explanation, mitigation or justification of this behaviour, EVS 2010 **fails** this evaluation criterion.

SENSITIVITY TO MIS-SPECIFICATION

I have already noted that EVS 2010 performs very well in the mis-specification tests in Suites 4 and 6. Suite 4 in particular, in which the magnitudes of error variances are mis-specified (in some cases dramatically so), does not provide any instances where the performance of EVS 2010 would give any concern at all. The only case that causes the method any serious trouble is test 6.2. This is the test in which the true A-distribution of flux shapes includes some values that are appreciably higher than any in the assumed flux shape distribution (and higher values of the flux shape parameter in these tests is associated with a lower ideal trip setpoint T).

Evaluation on Criterion 5 (sensitivity to mis-specification) – EVS 2010 has acceptable performance on the test suites involving mis-specification. As expected, the greatest sensitivity was observed when the assumed set of possible flux shapes excluded some true possibilities that were more extreme.

ANALYSIS OF RESULTS – EVS 2010 VERSUS BAYESIAN COMPARATOR

I now turn from absolute benchmarking of the EVS 2010 method against the criterion of a 95/95 tolerance limit to relative benchmarking of its performance against that of the Bayesian comparator, using all 6 evaluation criteria.

NON-COVERAGE

The Bayesian method generally has higher non-coverage values than EVS 2010, and in more instances this exceeds 5%. However, it has far fewer instances of non-coverage outside the range of 2.5% to 10%. Non-coverage in tests with no mis-specification exceeds 10% in 5 cases (out of 36), all with N = 20. It is below 2.5% also in just 5 cases. It therefore has a more balanced performance on this evaluation criterion than EVS 2010.

It is also noticeable that the non-coverage generally decreases with sample size (in marked contrast to EVS), and when N=500 it is below 5% for all tests in Suites 1, 2, 3 and 5.[10]

The Bayesian comparator does not appear to have the same difficulties with Suite 5. Whereas EVS non-coverage at N=500 is 7.44% for test 5.1 and 6.08% for test 5.2, the Bayesian values are respectively 3.45% and 4.05%. The performance of the Bayesian method in Suite 5 is similar to its behaviour in Suites 1, 2 and 3.

[10] It should perhaps be noted that the Bayesian method does not claim to be a 95/95 tolerance limit. It is the Bayesian analogue of a tolerance limit. The difference is subtle, however, and will not be pursued here.

MEAN DEFICIT

The mean deficit for the Bayesian method is generally a little larger than for EVS 2010, consistent with it having larger non-coverage, but similarly does not exceed 5% by much. There are only 3 instances when mean deficit exceeds 6% in Suites 1, 2, 3 and 5, all for N = 20. When N = 100 or 500, it matches EVS in being below 6% for all tests except in the mis-specified suites.

SD AND MEAN

The Bayesian comparator satisfies the basic face validity requirement that as sample size increases the SD decreases and the mean increases. However, mean and SD are also important measures for judging comparative efficiency, Evaluation Criterion 6.

There is a marked contrast between the methods in their SD values. Those for the Bayesian method are typically much smaller, particularly when N = 20 or 100. The fact that the Bayesian comparator obtains smaller SD values with the same sample sizes suggests that it is making better use of the evidence in the sample, eliminating more of the noise.

It is also striking that the Bayesian mean T* values are typically higher than those of EVS in almost every test. Since EVS generally has increasing means with increasing sample size, this is another indication that the Bayesian comparator is appearing to have a higher effective sample size than EVS.

A revealing test in this respect is test 2.1, which makes T not depend at all on the ripples parameter Q. Since the Bayesian method only uses the sample to learn about the distribution of Q, but that distribution is irrelevant in this test, we find that it gives exactly the same results for N = 20, 100 and 500. Non-coverage is also constant at 4%. The EVS method, in contrast, behaves in this test similarly to other tests – the mean increases with N and the SD decreases. The reason is that EVS uses the sample of S values not simply to learn about the A-distribution of Q. It also uses them in what the authors call the surrogate approach in order to compute various quantities such as the means and variances of aggregate error terms. So EVS computations have an additional component of noise. At N = 500, its results become the same as for the Bayesian method.

PARADOXES

If we examine the effect on the mean T* value for the Bayesian comparator in test Suite 3, as epistemic error variances are varied, we see that all the changes are in accord with intuition. The Bayesian method exhibits none of the paradoxical behaviour found in the EVS 2010 method when the epistemic error variances are varied and the assumed variances are correctly specified. This finding makes it hard to argue that the behaviour of EVS 2010 is natural and not counter-intuitive. Nevertheless, my assertion that EVS 2010 fails Criterion 4 remains provisional in case it can be argued either that its behaviour is appropriate or that the paradoxes will not arise in the NOP problem.

The behaviour in Suite 4 is different, but we should remember that intuition as to how the T* values should change is less clear in these tests. If we increase or decrease var(ε^{soro}) we find the mean T* value moves in the opposite direction to the changes seen in Suite 3. However varying the assumed var(ε^{rfsp}) value produces changes in the same direction as in Suite 3. This is different behaviour to that seen in the EVS 2010 method, although it is not entirely clear which, if any, of the results can be truly classed as paradoxical.

SENSITIVITY TO MIS-SPECIFICATION

The Bayesian method appears to be more sensitive to mis-specification. Whereas EVS only has trouble with test 6.2, the Bayesian method also has significantly raised non-coverage in tests 4.2, 4.4 and 6.4.

Suite 6 is quite revealing because the mean and SD of T* are the same in all these tests as in the base test (for each method and each sample size). What changes is the true $t_{1-\gamma}$ value. It is 1.379 in the base test and almost the same (1.378) in test 6.1. In test 6.1 the true distribution of φ is a continuous uniform distribution over [0,1], while the assumed distribution is a simple discretised version of that distribution. The two distributions are very similar, which explains why their true $t_{1-\gamma}$ values are almost the same. Accordingly, both methods give very similar non-coverage and mean deficit values to the base case. This turns out to be a very undemanding mis-specification and neither method shows any appreciable sensitivity. The test confirms the intuitive expectation that simply assuming a discrete distribution for φ is not in itself likely to damage performance.

In test 6.3 the true distribution of φ gives less probability to the higher values of φ than the assumed distribution. This leads to a higher value of $t_{1-\gamma}$, 1.464. Both methods quite naturally produce T* values that are too low and yield more or less zero non-coverage.

In tests 6.2 and 6.4 the opposite applies. These tests have lower $t_{1-\gamma}$ values, 1.142 and 1.33 respectively. The methods therefore have raised non-coverage values compared with the base test. The Bayesian comparator is more affected than EVS 2010 because of its higher mean. Arguably, EVS is less badly affected in terms of increased non-coverage because (a) it is less efficient (higher SD) and (b) even allowing for the higher SD its mean T* value is lower than it needs to be to achieve a non-coverage of no more than 5%.

SUMMARY

The primary focus of the relative benchmarking is to look at Evaluation Criterion 6, the relative efficiency of EVS 2010. The Bayesian comparator appears to be appreciably more efficient in its use of the available information.

Secondary interest lies in comparing the performance of EVS 2010 and the Bayesian comparator on the other 5 criteria. An important finding here is that the Bayesian comparator does not fail the face validity criterion concerning paradoxes. The Bayesian comparator may be deemed to perform better on criterion 1, concerning validity as a tolerance limit, which is assessed on the basis of non-coverage. It has non-coverage values that on the whole are more often close to the target 5%. On the other hand, EVS 2010's tendency to under-estimate leads to generally low non-coverage values, which may also be seen as desirable. Both methods perform very well on the mean deficit criterion, so that differences in non-coverage may not have significant practical implications. EVS 2010 appears to be more robust to mis-specification, again perhaps due to its tendency to under-estimate.

These findings should be interpreted with care because of the "quick and dirty" nature of the Bayesian comparator and the fact that it may be rather well suited to the particulars of the Group A scenario. They suggest that further development of the Bayesian approach may be worthwhile, but do not necessarily indicate inefficient performance on the part of EVS 2010.

Evaluation on Criterion 6 (efficiency) – EVS 2010 appears to be less efficient in the Group A benchmarking tests than the Bayesian comparator, although this may in part be due to the Bayesian comparator being particularly suited to the Group A scenario.

CONCLUSIONS AND RECOMMENDATIONS

My conclusions are organised into a series of comments, some of which draw together two or more of the evaluation findings. Each comment ends with recommendations. The comments (and associated recommendations) are categorised as major or minor, with the major comments presented first.

MAJOR COMMENTS

MAJOR COMMENT 1 – OVERALL PERFORMANCE

The overall performance of EVS 2010 in the benchmarking tests was good, although this finding is qualified by one major comment and some minor comments.

Performance on Evaluation Criteria 1 and 2, which relate directly to tolerance limit properties, indicate that EVS 2010 met these criteria with acceptable non-coverage and mean deficit in all those tests in which assumptions were not deliberately mis-specified, i.e. in Suites 1, 2, 3 and 5. However, further remarks and recommendations are given in Minor Comment 1.

Performance on Evaluation Criteria 3 and 4, which relate to face validity, was mixed. EVS 2010 passed the basic face validity test of Criterion 3 but provisionally failed the paradoxes test, Criterion 4. Major Comment 2 discusses this provisional failure and makes recommendations about how it might be addressed and resolved.

Performance on Evaluation Criterion 5, relating to sensitivity to mis-specification, indicates that EVS 2010 is relatively robust to the mis-specifications of assumptions in Suites 4 and 6. Some remarks and recommendations relating to the basis of this robustness and the importance of particular assumptions are made in Minor Comments 1, 2 and 3.

Performance on Evaluation Criterion 6, relating to efficiency, suggests that EVS 2010 makes less efficient use of available information than the Bayesian comparator, but this does not indicate that EVS 2010 is inefficient. The Bayesian comparator applies only to the simplified scenario of the Group A benchmark tests and has not been rigorously tested. See also Minor Comment 4.

MAJOR RECOMMENDATION 1: Subject to resolution of the face validity problem discussed in Major Comment 2 and Major Recommendation 2, and to any additional work deemed appropriate to address minor comments, EVS 2010 should be deemed to have passed the Group A Benchmarking Exercise.

MAJOR COMMENT 2 – PARADOX

Evaluation Criterion 4 set out a test of face validity that was motivated by previously expressed concerns about apparently paradoxical behaviour of EVS 2010 when variances of epistemic error terms were varied. It was not clear whether this arose only when the assumed variances were changed (implicitly keeping the true variances fixed by analysing the same data) or whether it could arise when both true and assumed variances were changed (i.e. with no mis-specification). A justification had been offered by AMEC NSS relating to the use of extremal operations in the T^0 function, but it was not clear how applicable this was and I had offered an alternative mechanism. Furthermore, a few isolated instances of paradoxical behaviour could have been due to chance.

Test Suite 3 has clarified this substantially. The paradoxical behaviour was found when the variance of the flux shape error term was varied, without mis-specification and without extremal operations. And it was found on average over thousands of simulated datasets. It was not found in Suite 3 when the ripples error variance was varied, but was found in Suite 4 when both error variances were varied (but Suite 4 involves mis-specification).

The intuition behind the face validity test seems unimpeachable to me when there is no mis-specification, and therefore the occurrence of paradoxical behaviour in Suite 3 is a serious cause for concern. Without face validity, it is hard to place trust in a method.

It remains possible that this behaviour can be mitigated through one of two kinds of reasoning. First, it could be shown that the intuition is faulty and that the apparently paradoxical behaviour is in fact natural and intrinsically sound. However, the fact that the Bayesian comparator does not suffer from this problem makes such reasoning less plausible. Second, it could be shown, through clear understanding of (i) how the behaviour arises and (ii) how more complex problems such as the NOP trip setpoint problem differ from the simplified Group A scenario, that this behaviour will not arise or will not cause difficulties in real applications.

MAJOR RECOMMENDATION 2: The developers of EVS 2010 should examine the causes of the paradoxical behaviour of EVS 2010 in Test Suite 3, with a view to justifying it or demonstrating that EVS 2010 will nevertheless behave acceptably in real applications for which it is proposed. Without satisfactory resolution of this problem, EVS 2010 should be deemed to have failed the Group A Benchmarking Exercise.

MINOR COMMENTS

MINOR COMMENT 1 – TOLERANCE LIMIT PROPERTIES

In applying the Evaluation Criteria 1 and 2, I raised some minor concerns about EVS 2010's performance in regard to the non-coverage and mean deficit outputs. I noted that the non-coverage was in most cases not within a factor 2 of the target 5%, being very often well below 2.5%. From one point of view this is perfectly acceptable because having non-coverage too low is at least erring on the safe side. However, it represents performance that is far from the target which the theory in the EVS 2010 Report says it should meet. So EVS 2010 is not performing as the theory indicates, and this is a source of concern because it suggests that other deviations from theoretical properties might be found in real applications.

It is useful in this context to remember the generally low reported values of mean deficit, which comes basically from the variance of T* being much lower than that of T, and is to be expected in other applications whenever appreciable quantities of data are available. As a result, even non-coverage that is above 5% will typically have low mean deficit. As pointed out in the discussion of mean deficit and in the analysis by AMEC NSS of the EVS 2010 outputs, even a non-coverage value as much as 8% to 10% can be seen as a minor deviation from the target. The case reported earlier in this report was the EVS 2010 non-coverage of 8% in test 4.5 with N = 500. Although this means that in this instance EVS 2010 has not performed as a valid 95/95 tolerance limit, and instead qualifies only as a 92/95 limit, the mean deficit is only 5.6% and AMEC NSS show that it can also be considered a 95/94.7 limit. The 8% non-coverage may be well above the target of 5% but if we simply relax the second 95% requirement very slightly to 94.7% the target is met. A β/γ tolerance limit is a package where both β and γ contribute. The true performance of a putative β/γ tolerance limit can be assessed by fixing γ and seeing whether the method achieves $100\beta\%$ coverage, or by fixing β and seeing whether the matching percentile on the T distribution is $100\gamma\%$. In the cited instance, it is equally valid to view the performance as 92/95 or 95/94.7.

The significance of the preceding remarks is that "being on the safe side" with a low non-compliance may be a positive feature, but having quite high non-compliance is also in effect relatively innocuous. I judge, therefore, that my concern that EVS 2010's low non-coverage indicates a deviation from the theoretical performance outweighs any comfort that low values are safe. EVS 2010's tendency to under-predict the true $t_{1-\gamma}$ is worthy of further investigation. It would be interesting to know whether there are any features of the implementation of the EVS 2010 theory that are consciously conservative. I am not aware of any, but suggest that the surrogate method might be a factor. The fact that the under-prediction is most marked when N is small supports this idea, since N also governs the amount of data available to the surrogate method.

In fact, it was also noted that the general increase of non-coverage with N might lead to it exceeding 5% routinely when N is sufficiently large. It would be unfortunate if one felt obliged to forgo additional data in case it caused poor performance. My expectation is that even if the downward bias were to disappear with sufficiently large N and non-coverage were to go above 5% it would still not be large enough to cause any concern, particularly in the light of earlier remarks.

MINOR RECOMMENDATIONS 1: CNSC should not judge proposed β/γ tolerance limit methods strictly on the basis of (non-)coverage for fixed γ. Even quite large exceedance of the nominal 100(1-γ)% non-coverage may correspond to perfectly adequate performance if we view β as fixed.

The fact that EVS 2010's non-coverage values are generally low in the Group A tests should be investigated to see if the cause might be found in some aspect of the way the theory is implemented.

It would be useful to extend some of the Group A tests to see whether non-coverage goes above 5% for sufficiently large N, and whether this is always accompanied by a small mean deficit.

MINOR COMMENT 2 – FLUX SHAPE DISTRIBUTION

Several interesting findings in the evaluation concern the uncertainty in the flux shape φ. The paradoxical behaviour of EVS 2010 that is the subject of Major Comment 2 arises particularly when we vary the variance of the errors in observations of the possible flux shapes φ_k; the only times when the EVS 2010's non-coverage exceeds 5% (excluding mis-specification tests) are when the weights attached to the possible flux shapes are varied from the uniform weighting of the base case; and the mis-specification in Suites 4 and 6 that most seriously affects EVS 2010 is allowing the true distribution of φ to extend beyond the assumed range. The second of these findings suggests looking in more detail at how EVS 2010 handles the flux shape uncertainty. The third is unsurprising, perhaps, but points to the importance of the assumption of a known set of possible flux shapes.

MINOR RECOMMENDATIONS 2: The investigation of the paradoxical behaviour of EVS 2010 (Major Recommendation 2) could usefully be extended to try to understand how T* is affected by non-uniform weights or shape in the A-distribution of φ.

It would be interesting to extend the Group A benchmark tests to include some where the number of possible flux shapes in much more than 20.

In any future application of EVS 2010, it is particularly important to be confident that there is no more than a very small chance of flux shapes arising more extreme than any in the set for which RFSP estimates have been obtained.

MINOR COMMENT 3 – MIS-SPECIFIED EPISTEMIC VARIANCES

Test Suite 4 produced some interesting results. In both EVS 2010 and the Bayesian comparator changes to the assumed epistemic error variances led to changes in the mean of T* that I will continue to call paradoxical, although the word is probably misplaced – the behaviour would only be genuinely paradoxical if the variance changes were made without mis-specification (as in Suite 3). The effects are probably readily explained in terms of the mis-specification but are nevertheless quite important to understand.

In some of their investigations, AMEC NSS have indicated that because of such behaviour it is better to under-estimate epistemic error variances than to over-estimate them when using EVS 2010. I feel that it would be dangerous to adopt this advice on the basis of a few numerical cases. I would prefer any such guidance to be based instead on deeper understanding of the behaviour. For, if we take the advice literally it would seem best to set the variances to zero, but it seems implausible that this would be a good thing to do.

MINOR RECOMMENDATION 3: The developers of EVS 2010 are invited to investigate more fully the behaviour of EVS 2010 when assumed epistemic error variances are changed, with a view to gaining enough understanding to make informed proposals on whether (and by how much) variances should in practice be deliberately under-estimated.

MINOR COMMENT 4 – EFFICIENCY

The apparently greater efficiency of the Bayesian comparator may not be an indication that EVS 2010 is unacceptably inefficient but it does suggest some scope for improvement. I had anticipated finding the Bayesian comparator to be more efficient because its inference route is more direct. Whereas the Bayesian comparator learns directly about the uncertain features of the A-distributions of Q and φ, EVS 2010 does so indirectly through data denoted by U or V that combine the two uncertainties. The less direct route also brings in the need to estimate some additional incidental quantities, for which the surrogate method is invoked.

I suspect that it is not feasible to develop a frequentist analogue of the more direct inference route of the Bayesian comparator. It seems to me, however, that there may be some scope to increase efficiency, and perhaps to reduce under-estimation, by better estimation of these incidental quantities. Perhaps the surrogate method could be refined by introducing shrinkage of the data, via a Bayesian or non-Bayesian shrinkage algorithm.

The Bayesian comparator as currently formulated is limited and simplistic, but its directness and simplicity are attractive and may genuinely be associated with greater efficiency. Further development may therefore be worthwhile, if only to provide a better comparative benchmark for EVS 2010.

MINOR RECOMMENDATIONS 4: The developers of EVS 2010 are invited to consider whether the apparent greater efficiency of the Bayesian comparator suggests possible improvements in EVS 2010.

Consideration should be given to further development of the Bayesian comparator, at least to provide a better comparator for EVS 2010 in more complex applications.

Report on the Benchmark B MCP Problem

Tony O'Hagan

November 19, 2012

1 Background

A major part of the CNSC contract 87055-10-1226 – R396.2: "Independent Verification and Benchmarking of Statistical Method and Mathematical Framework in OPG/BP 2010 EVS Methodology for Calculation of NOP Trip Setpoint" is benchmarking of the performance of the EVS method. The benchmarking exercise comprises two groups of tests. The Group A tests were based on a simplified, abstract problem which, although it formally fell within the remit of application of EVS was far from the primary focus of the contract, namely its application to calculation of the NOP trip setpoint. Nevertheless, the benchmark A testing revealed a number of interesting aspects of the performance of EVS that merited further exploration in the second group of tests. My report on the Group A testing is entitled "Report on EVS 2010 Group A Benchmarking Exercise", dated 22 February 2012.

The Group B tests are intended to examine performance in more realistic situations. The framework for these is set out in my report entitled "Proposal for Benchmark B problems, v3", dated 12 August 2012. It proposes two test problems; the first of these is referred to as the MCP problem. The Group B MCP problem is based on a test problem proposed by OPG/BP/AMEC, presented at a meeting with CNSC staff in March, 2012, and documented in AMEC report G0365/RP/001 R01 (June 19, 2012). The present report is an analysis of the results of the Group B MCP test suites.

2 The MCP test specification

The full specification of the MCP test suites for the Group B Benchmarking exercise is set out in my document "Specification for Benchmark B MCP problem, v1". A brief description is given here for reference.

2.1 The MCP problem

In the MCP problem we are interested in the maximum channel power

$$CP_{\max} = \max_{k=1}^{K} CP_k \ ,$$

where the true channel powers vector at any particular time is

$$\mathbf{CP} = \{CP_1, CP_2, \ldots, CP_K\}$$

and where $K = 480$ is the number of channels. We create a true A-distribution for $CP_{\max}$ by taking 1551 observed vectors and treating these as if they were true and define the entire set of possibilities. In this way we have a problem that is artificial but realistic. The true 95-th percentile of the 1551 $CP_{\max}$ values is the reference solution $t_{0.95}$.

Data for analysis by EVS 2010 or other methods are then created by simulating the effect of estimating the reactor state at a sample of points in time, using a suitable physics code. First N true vectors are sampled from the 1551 possible values (with replacement), where $N = 20, 100$ or 500. Then for $n = 1, 2, \ldots, N$ the n-th data vector is obtained by applying multiplicative channel-random and channel-common errors (E-errors) to these true states. The standard deviation of the common error is denoted by σ_0 and the standard deviation of random errors by σ_1. Using such data, EVS is applied to compute a 95/95 tolerance limit (i.e. an upper 95% confidence limit for $t_{0.95}$) W.

2.2 Relationship to the NOP problem

The MCP problem was discussed in the original EVS report, as a case of intermediate complexity before presenting the full NOP solution. It is simpler than the NOP problem in two important respects. The first is that there is no equivalent of the flux shapes; indeed, it is the treatment of flux shapes that follows the MCP problem in the EVS report to complete the full NOP analysis. Since some of the issues arising in the Group A benchmarking exercise (which did contain elements analogous to flux shapes), and also previously raised in my review of EVS, concerned the treatment of flux shapes, the MCP problem cannot be considered as an adequate realistic test case. This is why Group B also contains an NOP benchmarking problem. Nevertheless, the MCP problem is interesting partly because of its added simplicity. It allows us to investigate whether there are concerns about EVS that are not attributable to the treatment of flux shapes.

Another way in which the MCP problem is simpler is in the area of A-uncertainties. In the MCP problem the uncertainties relating to a future application (which mean that we are uncertain what the actual maximum channel power will be at any future time point) reside entirely in the distribution of the channel powers vector (which takes the role of the flux shapes in the EVS theory). In the benchmark version of the problem, this is completely determined by the 1551 possible vectors, all having equal probabilities in a future instance. The full NOP problem has some additional A-uncertainties; indeed, all of the uncertainties described as aleatory in the EVS report are absent from the MCP problem. This should in principle make it possible to obtain more accurate W values.

It is finally worth reiterating that the benchmark NOP problem is intended to be realistic but is not (and cannot be) real. By equating the A-distribution

RSP-0293

of the channel powers vector to 1551 observed vectors we achieve some degree of realism, but of course the true distribution of channel powers in a real application will differ from this benchmark case in several respects. First, it will not be discrete, and the complexity of the distribution of a genuine 480-dimensional quantity is barely described by any set of only 1551 instances. More important, the observed (or computed) vectors that we are assuming to be true realisations here have added observational errors. Realisations from the true distribution in a real application may be expected to be smoother. Nevertheless, the EVS theory does not make any assumptions of continuity or smoothness for the channel powers vector (or ripples vector), so despite these shortcomings of realism in the benchmark MCP problem it is still a valid test for EVS.

2.3 Alternative methods

In fact, four different methods were used to compute W values.

1. EVS
2. A Bayesian comparator, BC
3. A 'best estimate' method, BE
4. A 'traditional method', TM

Details of BC are given in the detailed specification document. It is intended as a comparator in much the same way as the Bayesian comparator for the Group A tests. That is, while it is not a genuine *competitor* (in the sense that it is expected to perform at least as well as EVS) because it has only been quickly put together for this exercise, its merit is that the underlying statistical theory is quite different from that of EVS. In the Group A testing, the comparator was useful because certain behaviours of EVS that gave cause for concern were not found in the comparator, so it could be concluded that such behaviours were not necessary or intrinsic to the problem. The role of BC in Group B tests will be similar, to check whether its behaviour is qualitatively the same as that of EVS in regard to the evaluation criteria.

The best estimate method, BC, simply computes the maximum channel power for each of the N simulated data vectors and reports the 95-th percentile of these values as W. Thus, this method treats the data as if they were true values, and it is acknowledged that such an approach does not properly account for errors in the data (which give rise to E-uncertainties regarding the underlying true values).

The traditional method is a version of the approach which I understand was used prior to the development of EVS (and is still in use by some companies). Although I have not studied the original SIMBRASS or ROVER codes, I am given to understand that their methods are based on the same approach. That is, the traditional method recognises the presence of errors in the data, and addresses that by simulating new vectors by adding channel-common and channel-random errors. The motivation for this approach is that the true vectors are equal to the observed data vectors minus the channel-common and channel-random errors, but since the distributions of the errors are symmetric, and are as likely to be positive as negative, adding and subtracting are equivalent.

The specification document identifies two versions of the traditional method to be used in the Group B tests. The specification document defined the 'true traditional method' to consist of taking the 98-th percentile of a large number of simulated $CP_{\max}$ values. The reason for using the 98-th percentile (which has been used, or at least proposed, in practice) is that there is just a single probability here, rather than the two 95% values in a 95/95 tolerance limit. The 98% point is therefore proposed to provide a degree of protection that is thought to be comparable to the tolerance limit. In addition to that method, described here as TM, the same analysis but using the 95-th percentile was performed and is referred to here as TTM (the 'true traditional method'). The second version of the traditional method, as set out in the specification document and referred to here as TM2, is a 95/95 variant intended to be closer in some sense to a tolerance limit method.

2.4 The MCP test suites

The MCP tests comprised a base case and two suites of variations. In each test case, three sample sizes were used, $N = 20, 100, 500$.

Base case

The base case is defined by setting both common and random error standard deviations to 1%, i.e. $\sigma_0 = \sigma_1 = 0.01$. The assumed values of both parameters equal their true values, so that there is no mis-specification.

Suite 1

Suite 1 comprises 8 cases. In all of these there is no mis-specification, so that the assumed standard deviations to be used in the methods are the true values.

1.1 $\sigma_0 = 0.01, \sigma_1 = 0.005$

1.2 $\sigma_0 = 0.01, \sigma_1 = 0.02$

1.3 $\sigma_0 = 0.005, \sigma_1 = 0.01$

1.4 $\sigma_0 = 0.02, \sigma_1 = 0.01$

1.5 $\sigma_0 = 0.005, \sigma_1 = 0.005$

1.6 $\sigma_0 = 0.02, \sigma_1 = 0.02$

1.7 $\sigma_0 = 0, \sigma_1 = 0.01$

1.8 $\sigma_0 = 0.01, \sigma_1 = 0.03$

The first 6 of these are simple variations around the assumed values. Case 1.7 links benchmarking Group B to earlier analyses performed by AMEC NSS and provides a code check. Case 1.8 allows one of the error standard deviations to go a little beyond the range of 0 to 2.5% that was stated as plausible at a meeting in Toronto.

Suite 2

This suite looks at mis-specification, with assumed error variances σ_0^a and σ_1^a deviating from true values σ_0 and σ_1 by ratios of 4:5 and 1:2. In each case the true values are as in the base case, i.e. $\sigma_0 = \sigma_1 = 0.01$.

2.1 Assume $\sigma_0^a = 0.01, \sigma_1^a = 0.005$

2.2 Assume $\sigma_0^a = 0.01, \sigma_1^a = 0.008$

2.3 Assume $\sigma_0^a = 0.01, \sigma_1^a = 0.0125$

2.4 Assume $\sigma_0^a = 0.01, \sigma_1^a = 0.02$

2.5 Assume $\sigma_0^a = \sigma_1^a = 0.005$

2.6 Assume $\sigma_0^a = \sigma_1^a = 0.008$

2.7 Assume $\sigma_0^a = \sigma_1^a = 0.0125$

2.8 Assume $\sigma_0^a = \sigma_1^a = 0.02$

2.5 Performance measures and evaluation criteria

The performance measures are equivalent to those employed in the Group A tests.

- *Mean.* The mean of the M solutions.
- *SD.* The standard deviation of the M solutions.
- *Non-coverage.* The proportion of the M solutions that lie below $t_{.95}$. Ideally, this should be 5%.
- *Mean deficit.* For every $W^{[m]}$ that is below $t_{.95}$, the deficit is the proportion of $CP_{\max}^{(d)}$ values that lie above $W^{[m]}$. The mean deficit is the average of the deficit values for all $W^{[m]}$ below $t_{.95}$. By convention if no values are below $t_{.95}$ then the mean deficit is 0.05.

Several evaluation criteria were defined for the Group A tests, and versions of these are employed in the MCP problem. The criteria are also informed by issues revealed in the Group A testing.

Tolerance limit criteria

The first criterion concerns the non-coverage measure. This should ideally be no higher than 5%. Theoretically, it should be 5% in all cases where there is no mis-specification, although the practicalities of implementation mean that this will not be achieved exactly. Where there is minor mis-specification, of an order of magnitude that could readily arise in practice, the non-coverage should not rise much above 5%.

- *Criterion 1 (Desirable).* The non-coverage should be 5% or less in the base case and all tests in Suite 1. It should also be less than, or not much more than, 5% in Suite 2 tests where the mis-specification is minor (which will here be interpreted as tests 2.2, 2.3, 2.6 and 2.7 but may be open to discussion).

The second criterion makes use of the mean deficit measure to moderate the non-coverage. Group A testing found that even in tests where non-coverage exceeded 5% the mean deficit was typically only slightly above 5%. Moderately raised non-coverage can be tolerated when mean deficit is not high. We can also excuse raised non-coverage when $N = 20$ because EVS is not intended to be used with such small samples. Criterion 2 demands that deviation from the strict conditions of Criterion 1 'must not be excessive'. The precise interpretation of what is or is not excessive is a matter for the industry and its regulator to determine, but for the purposes of the analysis in this report I have chosen limits which seem plausible to me.

- *Criterion 2 (Essential).* In all tests to which Criterion 1 applies, for both $N = 100$ and $N = 500$, non-coverage must not be excessive (which will here be interpreted as not larger than 10% but is open to discussion) *and* when non-coverage exceeds the levels set out in Criterion 1 the mean deficit must not be excessive (interpreted here as not larger than 15% but also open to discussion).

The third criterion relating to the tolerance limit property is concerned with the efficiency of a method. The first two criteria can be achieved by a method that is very conservative, and so produces W values that are generally unnecessarily high. For instance, a method that set W to 1000 times the largest value found in any of the N sample data vectors would comfortably satisfy the first two criteria but would be ridiculous in practice. So it is desirable for a method not to be excessively conservative.

- *Criterion 3 (Desirable).* In the base case and all of the tests in Suite 1, the non-coverage should be closer to 5% than to 0%.

Face validity criteria

Group A benchmark tests revealed some issues over face validity of EVS, and so it is important for Group B tests to apply similar criteria. These criteria concern the fact that as the quantity or accuracy of data increases so any statistical method should be able to make better inferences. In the case of an upper tolerance limit (as is applied in the MCP problem), this should be reflected in a reduced mean and reduced SD.

- *Criterion 4 (Essential).* As the sample size N increases in any test, the mean and SD must both decrease.

The final two face validity criteria apply *only* when there is no mis-specification, i.e. to the base case and Suite 1.

- *Criterion 5 (Desirable).* If either σ_0 or σ_1 increases, then the mean and SD should both increase.

Criterion 5 has been made just 'desirable' rather than 'essential' because of the face validity problems that arose for EVS in the Group A tests. There were instances where increasing E-error variances led to a mean that did not decrease (as it should have done for a lower tolerance limit) but a SD that increased (as it should do for either kind of tolerance limit). The implication was that for sufficiently large E-error variances large non-coverage values could arise. It was accepted that failure of the face validity in those tests was a real feature of EVS but was argued to be a small effect that would not cause excessive non-coverage in realistic applications. So Criterion 5 is only set to 'desirable' but another, somewhat weaker, criterion is 'essential'.

- *Criterion 6 (Essential).* If Criterion 5 is not met for any comparison of tests with $N = 100$ or $N = 500$, then it must be clear that for all reasonable values of σ_0 and σ_1 the non-coverage (and mean deficit) will remain acceptable.

As with Criterion 2, Criterion 6 uses terms, 'reasonable' and 'acceptable', the precise meanings of which are matters for the industry and (primarily) its regulator to determine. Ultimately, the question of fitness for purpose necessarily involves judgements like these.

3 Results and evaluation

The values of the mean, SD, non-coverage and mean deficit measures for each of the 6 methods in each of the 17 tests and the 3 sample sizes are given in Tables 1 to 4 respectively.

3.1 Tolerance limit criteria — EVS

The non-coverage and mean deficit measures in Tables 3 and 4 determine performance of the various methods against the tolerance limit evaluation criteria.

Criterion 1

> "The non-coverage should be 5% or less in the base case and all tests in Suite 1. It should also be less than, or not much more than, 5% in Suite 2 tests where the mis-specification is minor (which will here be interpreted as tests 2.2, 2.3, 2.6 and 2.7 but may be open to discussion)."

We see from Table 3 that the condition is satisfied in the majority of instances, but not all. Most exceptions, however, arise when $N = 20$ and these give no real cause for concern. AMEC NSS have consistently said that $N = 20$ is too small a sample for EVS to be expected to work well.

There are nevertheless several instances where $N = 100$ or $N = 500$ and yet non-coverage is larger than 5%. Consider first the tests where there is no mis-specification, i.e. the base case and Suite 1.

- When $N = 100$, non-coverage is 8.7% in test 1.1, 14.2% in test 1.4, 6.3% in test 1.5 and 10.9% in test 1.6. There seems to be no real pattern to these instances. Tests 1.1 and 1.5 have the smallest value of σ_1 coupled with the two smallest values of σ_0, while tests 1.4 and 1.6 have the highest value of σ_0. Although the exceedances over 5% might be due to sampling variation in some of these tests, this cannot be the explanation in all cases (particularly in test 1.4).

- When $N = 500$, non-coverage is 6.7% in test 1.1. This is the only instance with the largest sample size, but although 6.7% is not far above 5% the exceedance cannot realistically be explained by sampling variation. (Formally, 6.7% is 5 standard errors above 5%.)

Moving to consideration of the tests involving mild mis-specification, the criterion only asks for non-coverage to be 'not much more than 5%'. However, whilst this is clearly achieved in tests 2.2 and 2.6, the situation is quite different in tests 2.3 and 2.7.

- In test 2.3, non-coverage is 11.6% when $N = 100$ and 16.2% when $N = 500$. In test 2.7, it is 42.7% when $N = 100$ and reaches 95.5% for $N = 500$. There is a clear pattern here, which is that assuming error variances smaller than the true values is what causes the problem, and the problem actually gets worse for $N = 500$. Whatever the interpretation of 'not much more than 5%', these values surely do not meet that requirement.

AMEC NSS have also consistently said that it is important not to under-estimate the E-error variances, and this is firmly supported by these test findings. The important point is that we see markedly deteriorating performance when the mis-specification is still modest, just by a factor of 1.25. Even when consciously attempting to err on the high side with these variances, the difficulties of obtaining reliable, high-quality data from which to estimate them mean that it could still be possible to under-estimate.

RSP-0293

The conclusion must be that EVS has not satisfied Criterion 1. However, this criterion is specified only as desirable, not essential.

Criterion 2

> "In all tests to which Criterion 1 applies, for both $N = 100$ and $N = 500$, non-coverage must not be excessive (which will here be interpreted as not larger than 10% but is open to discussion) *and* when non-coverage exceeds the levels set out in Criterion 1 the mean deficit must not be excessive (interpreted here as not larger than 15% but also open to discussion)."

Criterion 2 provides the weaker but essential counterpart to Criterion 1. We have seen that in some of the tests to which Criterion 1 applies do have non-coverage over 10%. It is arguably never excessive in the tests without mis-specification (allowing for sampling variation), but it can clearly be excessive when there is mild mis-specification.

A notable feature of the results in the MCP benchmark test is the high mean deficit values that have arisen in several instances. In Group A tests, mean deficit was only slightly above 5% except in cases of substantial mis-specification when non-coverage was also high. We see here instances where non-coverage is small yet mean deficit is surprisingly high – for instance in test 1.2 with $N = 100$ non-coverage is 0.2%, meaning that only two of 1000 sample runs gave a W value below the true $t_{0.95}$, yet these two values were obviously well below $t_{0.95}$ because the mean deficit is 16.5%! This points to a high degree of skewness in the sampling distribution of W which may be a cause for concern if repeated in the NOP problem.

Nevertheless, high mean deficit is not a problem as long as non-coverage is not also raised. We therefore consider the mean deficit in the cases noted in bullet points in the discussion of Criterion 1. We see just the following cases for concern.

- In test 1.4 with $N = 100$, where non-coverage is 14.2%, mean deficit is 29.3%.
- In test 2.7 with $N = 100$, where non-coverage is 42.7%, mean deficit is 30.8%.

These values indicate a failure of Criterion 2 at $N = 100$. It seems that EVS really needs a sample size nearer to $N = 500$ (where we see no clear violations of Criterion 2) in order to be reliable when there is no mis-specification. Even with $N = 500$, non-coverage can be excessive when there is relatively mild under-estimation of E-error variances.

Criterion 3

> "In the base case and all of the tests in Suite 1, the non-coverage should be closer to 5% than to 0%."

RSP-0293

This is rarely the case for EVS, which exhibits non-coverage values equal to zero in most of the tests with $N = 500$. It seems that EVS generally errs on the 'safe side' in the MCP benchmark tests, just as it did in the Group A tests. However, whilst this criterion is desirable it is not essential.

3.2 Tolerance limit criteria — other methods

Non-coverage is very small for the other 5 methods. In general, we expect on theoretical grounds that TM, TTM, TM2 and BE will over-estimate $t_{0.95}$ (in the MCP problem where we are looking for an upper limit). This is borne out in the tests, where non-coverage is zero for TM, TTM and TM2 in every instance (except one case where TTM is 0.2%). Non-coverage for BE is also zero or very close to zero except in one notable instance; in test 1.5 with $N = 20$ its non-coverage is 13.7%. Test 1.5 has the smallest E-error variances, so that the theoretical tendency for BE to over-estimate is minimal, and with $N = 20$ there is sufficient sampling variation to mean it under-estimates some of the time.

The tests also show that BC consistently over-estimates. Apart from three instances (non-coverages of 0.07%, 0.03% and 0.03%) non-coverage is again zero throughout. This is not for any theoretical reasons but apparently because the Bayesian modelling is too simplistic for this problem.

All of these methods clearly satisfy Criteria 1 and 2, but perform appreciably worse than EVS on Criterion 3.

3.3 Face validity criteria — EVS

The mean and SD performance measures, given in Tables 1 and 2 respectively, are concerned with performance of the various methods against the tolerance limit evaluation criteria.

Criterion 4

> "As the sample size N increases in any test, the mean and SD must both decrease."

EVS passes this essential criterion in every test.

Criterion 5

> "If either σ_0 or σ_1 increases, then the mean and SD should both increase."

Remember that the face validity criteria 5 and 6 are considered only for the base case and Suite 1. We have three separate sequences of tests in this group in which the random error standard deviation σ_1 increases while σ_0 is constant. These are 1.5→1.3 ($\sigma_0 = 0.005$); 1.1 →Base→1.2→1.8 ($\sigma_0 = 0.01$); 1.4→1.6 ($\sigma_0 = 0.02$). The arrows indicate increasing σ_1. In these sequences there are 5 separate cases of increasing σ_1 to which Criterion 5 applies. EVS satisfies Criterion 5 fully (apart from one instance where the SD decreases when $N = 20$, but this can be explained by random variation).

There are also three sequences in which σ_0 increases while σ_1 is constant. They are 1.5→1.1 ($\sigma_1 = 0.005$); 1.7→1.3→Base→1.4 ($\sigma_1 = 0.01$); 1.2→1.6 ($\sigma_1 = 0.02$). There are again 5 separate cases of increasing σ_0 to which Criterion 5 applies. The performance of EVS in these sequences is less consistent. The SD measure does increase as required but the mean does not always; in 3 of the 5 comparisons the mean actually decreases when $N = 500$. Some changes of mean are small as σ_0 increases, and may be explainable by sampling variation. However, from test 1.2 to test 1.6 the mean decreases by 4.5, which is not explainable in this way (formally, 4.5 is almost 5 standard errors). Conversely, from the base case to test 1.4 the mean correctly increases by an amount, 7.0, which also cannot conceivably be explained by random variation (more than 8 standard errors).

The contrary movement of the mean when the channel-common error standard deviation σ_0 varies is the same kind of face validity failure as was observed in the Group A benchmarking exercise, and was described there as paradoxical and undesirable. However, it was not expected to arise in the MCP problem because the only E-uncertainties here are analogous to the error in the ripples term in the NOP problem, whereas no paradoxical behaviour was found with changes in the ripples error variance in Group A. The Group A face validity failure was in respect of changes to the flux shape error variance, and there is no flux shape equivalent in the MCP problem.

So it appears that paradoxical failures of face validity are more pervasive in the EVS applications than had been thought on the basis of previous testing. Perhaps the most serious aspect of this is that the paradoxical behaviour arises here when $N = 500$ and is absent (or less apparent) for smaller samples.

However, Criterion 5 is specified as desirable, rather than essential.

Criterion 6

> "If Criterion 5 is not met for any comparison of tests with $N = 100$ or $N = 500$, then it must be clear that for all reasonable values of σ_0 and σ_1 the non-coverage (and mean deficit) will remain acceptable."

This criterion is intended to provide a test for whether paradoxical behaviour, whilst always undesirable, would actually matter in practice. The issue is that if the SD of W is increasing while the mean is decreasing (or even staying constant), then unless the shape of the distribution of W is also changing in a compensatory way (for which there is no evidence, and indeed there is some evidence to the contrary) the non-coverage will increase (and also mean deficit). However, this only becomes a problem if the non-coverage can become unacceptably high without the relevant E-error variance having to be made unrealistically large. Criterion 6 accepts that paradoxical behaviour may happen but asks whether we can be confident that for all realistic variances the non-coverage will not be unacceptable.

Test Suite 1 was planned principally to study this kind of behaviour in the long sequences containing three successive comparisons to which the criterion could be applied. However, the long sequence in which σ_0 increases (1.7→1.3→Base→1.4) does not show any consistent paradoxical movements. We have decreases in mean initially but an emphatic increase at the end. This certainly does not suggest that problems would arise at sufficiently large σ_0; on the contrary the decreases disappear when σ_0 becomes sufficiently large. The most striking example of a decreasing mean (1.2→1.6) arises when both σ_0 and σ_1 are already relatively large, and again it is not easy to imagine non-coverage increasing sufficiently without moving σ_0 to a very unrealistically high value.

Overall, EVS satisfies Criterion 6 as far as it is possible to ascertain from limited testing. The principal remaining concern here is that the occurrence of paradoxical movements seems to be more pronounced with increasing N, raising the question of whether Criterion 6 might fail if N were appreciably greater than 500.

3.4 Face validity criteria — other methods

BC and TM2 both pass Criterion 4 in every relevant test. However, TM, TTM and BE, while satisfying the condition for the SD measure to decrease as N increases, fail on the requirement for the mean also to decrease. In effect, the mean is independent of N for these methods, any variation in the data being due to sampling. This is because these are not true tolerance limit methods.

All of the alternative methods satisfy Criterion 5 in respect of the requirement for the mean to increase as error variances increase. The paradoxical movements of the mean that are seen with EVS are not exhibited by any of these other methods. In particular, the comparison between EVS and BC demonstrates that paradoxical behaviour of EVS is not a necessary and intrinsic feature of the problem.

However, it is noticeable that the requirement for the SD also to increase as error variances increase (which was satisfied by EVS) is not satisfied by BC, TM and TTM. For TM and TTM, this seems to be a property of the maximum operation. It is more surprising to see this paradoxical behaviour in BC, which is intended to be a proper Bayesian tolerance interval method. Although Criterion 5 is failed by these 3 methods, all 5 automatically satisfy Criterion 6 (because increasing mean but decreasing variance is the opposite problem to that experienced by EVS).

4 Conclusions

The principal findings of the MCP benchmarking exercise concerning the performance of EVS in relation to the evaluation criteria are presented and briefly discussed in the following subsections.

4.1 Tolerance limit performance

In regard to the tolerance limit criteria, EVS performs generally well in tests where there is no mis-specification and with the larger sample sizes $N = 100$ and $N = 500$. As has been acknowledged by AMEC NSS, performance is not adequate for $N = 20$, and there is a suggestion in test 1.4 that even $N = 100$ may not always be adequate to ensure satisfactory tolerance limit performance. Unlike the Group A benchmark problem, the MCP problem has produced large mean deficit values, indicating heavy tails in the distribution of W. This increases the importance of achieving satisfactory coverage.

> **Principal finding 1.** In the MCP benchmark tests, EVS performance in regard to the tolerance limit Criteria 1 and 2 is found to be poor for low sample sizes. $N = 100$ does not appear to be large enough to satisfy these criteria consistently, even when error variances are correctly specified.

EVS has shown a high degree of sensitivity to mis-specification of the E-error variances σ_0 and σ_1. Even modest overstatement of these parameters, which has been taken here to mean by a factor of 1.25, leads to values for the non-coverage and mean deficit measures that are judged to be unsatisfactorily high. Unless it can be shown that in practical applications of EVS over-estimation by anything approaching this amount will not arise, EVS must be deemed to have failed the essential Criterion 2. It is particularly important to note that this sensitivity is higher at $N = 500$ than at $N = 100$. Whereas larger N seems to ensure good performance when variances are correctly specified, it seems that larger N may make things worse when they might be mis-specified (which in practice will always be a risk).

In general when variances are not over-estimated, EVS tends to over-estimate $t_{0.95}$ and for the larger sample sizes its non-coverage thereby tends to be well below the nominal 5%. Although this means that it does not satisfy the desirable Criterion 3, it should be noted that EVS does much better in this regard than any of the alternative methods examined in this exercise. That is does better than the ‘traditional’ and ‘best estimate’ methods is to be expected, but it also easily out-performs the Bayesian comparator in this respect. It may be possible to develop a more efficient method, i.e. one that satisfies the criteria without such a ‘conservative’ tendency, but it is clear that to do so will not be a simple challenge.

> **Principal finding 2.** In the MCP benchmark tests, EVS performance on the tolerance limit Criteria 1 and 2 is highly sensitive to mis-specification of error variances. It fails Criteria 1 and 2 when error variances are overstated by as little as a factor of 1.25. Performance becomes worse with increasing sample size. This is a serious concern, because in practice it will not be easy to ensure that error variances are not over-estimated.

4.2 Face validity performance

EVS behaviour that was described in the Group A report as paradoxical and undesirable has been seen again in the MCP benchmark tests. It should be noted that in both benchmarking exercises the examples which fail the face validity test do so only in minor ways, and formally EVS has passed the essential Criterion 6. Nevertheless, it seems that the existence of this kind of behaviour in EVS is more widespread than had been thought on the basis of the Group A analyses, because in the MCP problem it arises when varying E-error parameters that play a quite different role in the theory to the ones that caused paradoxical behaviour in Group A. Another potential cause for concern is the fact that the contrary mean shifts may be more marked when N is larger.

> **Principal finding 3.** In the MCP benchmark tests, EVS exhibits paradoxical behaviour which fails Criterion 5. Although I deem that this behaviour is not sufficiently marked for EVS to fail the essential Criterion 6, there remain two causes for concern. First, the emergence of paradoxical movements of mean values when error variances are change (without being mis-specified) appears to be a widespread and intrinsic problem with EVS. Second, although these effects have generally been found to be small, there is evidence that they become more marked with increasing sample size.

RSP-0293

Table 1 **Mean**

	EVS			BC			TM			TTM			TM2			BE		
Test	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500
Base	6918	6872	6851	6987	6957	6929	7061	7065	7066	7010	7012	7012	7092	7051	7030	6928	6927	6927
1.1	6897	6850	6833	6910	6888	6866	6979	6983	6984	6933	6935	6936	7003	6970	6952	6883	6883	6884
1.2	7013	6951	6922	7136	7093	7054	7301	7304	7305	7234	7235	7236	7350	7289	7260	7063	7062	7061
1.3	6913	6870	6851	6979	6953	6927	6998	7001	7002	6957	6960	6960	7022	6990	6974	6896	6896	6895
1.4	6956	6891	6858	7015	6968	6933	7230	7237	7238	7147	7151	7152	7281	7215	7181	7017	7018	7018
1.5	6884	6847	6832	6899	6884	6865	6908	6912	6913	6877	6878	6878	6922	6901	6889	6850	6849	6850
1.6	7030	6954	6918	7157	7101	7058	7450	7455	7457	7357	7359	7361	7511	7433	7394	7145	7140	7140
1.7	6912	6871	6852	6975	6951	6926	6973	6976	6977	6937	6939	6938	6995	6966	6951	6884	6883	6883
1.8	7136	7049	7014	7236	7184	7139	7596	7599	7599	7508	7509	7509	7665	7580	7541	7234	7228	7229
2.1	6947	6900	6881	6958	6937	6912	7021	7026	7027	6974	6975	6976	7045	7011	6992	6929	6926	6927
2.2	6930	6885	6866	6976	6948	6922	7042	7047	7048	6992	6995	6996	7070	7032	7013	6926	6926	6928
2.3	6895	6851	6830	7000	6967	6937	7088	7091	7092	7034	7037	7038	7122	7078	7056	6925	6927	6928
2.4	6813	6766	6742	7020	6982	6949	7191	7194	7194	7130	7132	7133	7235	7181	7155	6927	6927	6927
2.5	6997	6942	6921	6985	6969	6943	6987	6994	6995	6947	6950	6950	6997	6976	6963	6927	6926	6927
2.6	6958	6907	6885	6987	6963	6937	7027	7032	7033	6982	6984	6984	7051	7018	7000	6928	6928	6927
2.7	6857	6810	6790	6995	6964	6936	7109	7111	7112	7051	7053	7054	7147	7098	7074	6927	6927	6927
2.8	6674	6649	6639	7022	6982	6949	7278	7280	7280	7200	7203	7203	7334	7265	7232	6924	6927	6927

RSP-0293

Table 2 **SD**

	EVS			BC			TM			TTM			TM2			BE		
Test	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500
Base	87.0	29.1	12.2	34.5	14.6	5.8	34.6	15.3	7.0	27.6	13.0	5.8	33.8	13.7	5.9	40.7	20.1	9.4
1.1	216.7	24.2	10.5	36.3	16.1	6.4	35.3	15.8	7.2	28.4	12.9	6.1	33.6	13.9	6.1	37.8	19.0	8.9
1.2	125.2	41.2	17.2	32.6	13.2	5.3	35.0	15.6	7.0	28.8	13.5	5.8	36.2	14.2	6.1	51.4	25.8	11.7
1.3	65.3	24.9	9.9	30.5	12.7	4.8	28.7	12.8	5.9	23.0	10.7	4.8	27.7	11.2	4.8	35.8	18.1	8.0
1.4	156.4	69.5	21.9	57.0	25.0	10.5	54.5	24.1	10.9	46.5	21.6	9.1	53.9	21.8	9.5	62.5	31.7	14.2
1.5	52.0	19.9	8.0	31.5	14.2	5.5	29.6	13.6	6.2	24.4	11.3	5.0	27.1	11.5	5.0	31.0	15.9	7.1
1.6	200.2	80.8	23.8	51.1	21.8	9.2	53.8	23.9	10.8	49.9	21.6	9.4	54.3	21.9	9.5	74.9	35.0	15.2
1.7	62.3	22.1	9.1	29.3	12.2	4.5	27.0	12.2	5.6	22.8	9.6	4.2	25.9	10.5	4.5	34.3	16.0	7.3
1.8	174.6	57.7	23.3	32.5	12.9	5.2	38.5	17.2	7.7	31.2	14.1	6.6	41.4	15.7	6.5	65.6	32.2	14.5
2.1	79.5	25.9	10.9	42.2	18.8	7.2	39.3	17.5	8.0	33.6	14.2	6.4	36.6	15.1	6.6	43.7	20.9	9.2
2.2	86.4	26.9	11.7	37.6	16.1	6.3	36.5	16.2	7.4	29.8	13.1	6.0	35.0	14.2	6.2	42.0	20.0	9.4
2.3	94.0	29.7	13.2	31.0	13.1	5.3	32.5	14.3	6.5	27.1	12.3	5.7	32.3	13.0	5.6	43.4	20.2	9.3
2.4	116.9	41.6	15.4	25.5	11.0	4.6	27.8	12.4	5.6	24.7	11.2	5.1	28.6	11.6	5.1	41.7	21.2	9.3
2.5	70.0	25.5	10.6	45.5	19.7	7.5	45.8	20.8	9.5	37.3	16.5	7.2	39.5	16.9	7.4	43.8	20.8	9.5
2.6	76.4	27.3	11.0	39.0	16.4	6.4	38.6	17.1	7.8	31.8	14.5	6.3	36.1	14.8	6.4	43.5	20.8	9.5
2.7	120.7	67.0	21.1	29.5	12.5	5.0	30.8	13.6	6.2	26.1	12.1	5.3	31.3	12.5	5.4	42.6	20.8	9.3
2.8	48.8	11.8	4.9	23.6	10.2	4.3	24.8	11.0	5.0	23.6	10.6	4.7	26.1	10.7	4.7	40.7	20.3	9.4

RSP-0293

Table 3 **Non-coverage**

	EVS			BC			TM			TTM			TM2			BE		
Test	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500
Base	13.7	2.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
1.1	12.9	8.7	6.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0
1.2	10.8	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.3	4.5	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0
1.4	28.1	14.2	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.5	8.3	6.3	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	13.7	1.2	0.0
1.6	19.6	10.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.7	3.7	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
1.8	3.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2.1	7.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2.2	9.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
2.3	17.2	11.6	16.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
2.4	46.2	92.6	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0
2.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
2.6	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
2.7	41.6	42.7	95.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
2.8	98.6	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0

RSP-0293

Table 4		Mean deficit																
		EVS			BC			TM			TTM			TM2			BE	
Test	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500	N = 20	N = 100	N = 500
Base	25.7	12.3	5.7	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	8.2	5.0	5.0
1.1	26.3	6.9	5.8	7.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	7.6	5.0	5.0
1.2	15.4	16.5	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
1.3	17.7	6.6	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.4	5.0	5.0
1.4	23.4	29.3	25.1	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
1.5	12.9	6.1	5.5	6.0	5.0	5.0	5.0	5.0	5.0	9.2	5.0	5.0	5.0	5.0	5.0	8.1	6.0	5.0
1.6	14.9	13.2	18.3	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
1.7	12.1	6.4	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.9	5.0	5.0
1.8	9.3	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
2.1	17.8	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
2.2	21.9	6.6	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.5	5.0	5.0
2.3	35.6	9.3	6.3	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	7.5	5.0	5.0
2.4	56.1	26.3	35.1	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.3	5.0	5.0
2.5	12.8	5.0	5.0	6.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	12.8	5.0	5.0
2.6	18.6	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
2.7	36.9	30.8	13.3	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	8.1	5.0	5.0
2.8	80.6	90.3	93.8	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.2	5.0	5.0

Report on NOP Benchmark Tests

Anthony O'Hagan

February 10, 2013

The full specification of the Group B NOP benchmarking tests is set out in my report 'Specification for Benchmark B NOP problem, v2', dated 7th November, 2012. It makes use of data provided in the appendices to a letter from AMEC NSS entitled "Re: Compilation of input data to be used for Group B NOP benchmarking exercises", dated 4th October, 2012 and bearing the document number G0365/039/000001 R02. This will be referred to herein as the technical document. The results of the tests, in the form of the values of the specified performance measures, are given in the attachment to a letter from AMEC NSS dated 20th December, 2012, and bearing their reference number G0365/051/000002 R00. Because of an error in the original specification, some additional tests were run, and the full data including these additional tests may be found in the Appendix to this report.

This report reiterates the key details of the test specification and contains the conclusions of my analysis of those results against the six Group B benchmarking criteria.

1 The NOP test scenario

The NOP test problem closely follows the mathematical specification in the updated EVS 2010 document (G0263/RP/008 R01, dated October 4, 2011). All equation numbers herein refer to equations in that document.

1.1 A-distributions and the target trip setpoint

First consider the uncertainties in an actual instance where the trip setpoint is applied. These are A-uncertainties, characterised by A-distributions. The EVS terminology is 'aleatory' uncertainties and distributions. Note that I use the term 'uncertainty' always to refer to the degree to which a quantity of interest is unknown. A-uncertainties arise because the true detailed reactor state at a time when the trip setpoint is applied is unknown. There are three kinds of A-distributions.

1. Ripples. In the benchmark tests, the ripples vector Q will be assumed to have a discrete A-distribution taking 1551 equi-probable values. (As in the MCP benchmark tests, these Q vectors are obtained from observed

data but are treated as true values for the purposes of the test suite.) The channel powers CP are then obtained by multiplying Q by the reference channel powers as in equation (6), with the reactor power RP set to 100%.

2. Flux shape. In the base case, the flux shape φ will be assumed to have a discrete A-distribution taking 70 possible values corresponding to flux shape category MCA (shapes 9 to 78) in Appendix B of the technical document, and with probabilities (or weights) as given in that Appendix. The corresponding true values of the channel overpowers COP, flux overpowers FOP and reference critical channel powers CCP^0 vectors are all assumed to be the values given by the relevant physics codes RFSP and TUF. (As in the MCP benchmarking, this again treats 'observations' as if they were true values, purely for the purpose of generating a benchmark test suite.)

3. A-error terms. The values of the indicated reactor power RP^{ind}, actual critical channel powers CCP and fluxes at detector locations FX are then given by applying random A-errors ε^{rp}, θ^{ccp} and θ^{dr} as in equations (10) to (12). In equation (12), dr is set to 1. The A-distributions of ε^{rp}, θ^{ccp} and θ^{dr} are assumed to be normal, with means and standard deviations set as in Appendix A of the technical document. These values are repeated for convenience in Table 1 below. To clarify the precise meaning of common and specific errors, consider the term θ^{ccp}. CCP is a vector whose j-th value CCP_j is the critical channel power for channel j. Equation (11) defines $CCP_j = CCP_j^0(1 + \theta_j^{ccp})$, where θ_j^{ccp} is the sum of a common error δ^{ccp}, which is normally distributed with mean zero and standard deviation 0.00476, and a specific error η_j^{ccp} which is normally distributed with mean zero and standard deviation 0.00161. The common error and all the specific errors are independent. The calibration drift error has no common error term. The error in reactor power is a single number rather than a vector, so it has no specific error term; it also has a mean of −0.015.

	Specific error SD	Common error SD	Error mean
Calibration drift error, θ^{dr}	0.012	0.0	0.0
Error in reactor power calculations, ε^{rp}	–	0.005	−0.015
A-error in critical channel power, θ^{ccp}	0.00161	0.00476	0.0

Table 1. Standard deviations (SD) and means of A-distributions for error terms

Computing the target trip setpoint

The true, or ideal, trip setpoint is then computed using equation (5) with the calibration factor CF set to 1.1. This is subject to A-uncertainties and so has its A-distribution which is evaluated empirically by Monte Carlo sampling. Thus, a random Q is selected from the discrete A-distribution in item 1 above and

CP is thereby implicitly sampled; a random φ is selected from the discrete A-distribution in item 2 above and the corresponding values of COP, FOP and CCP^0 are thereby implicitly sampled; random values of ε^{rp}, θ^{ccp} and θ^{dr} are sampled from their normal distributions as in item 3 above and Table 1 and random values of RP^{ind}, CCP and FX are thereby implicitly sampled; equation (5) is applied to obtain a sampled value of TSP; and finally this is all repeated many times to obtain many Monte Carlo values for TSP.

The target ('true') trip setpoint $t_{0.95}$ is the lower 95% percentile of the A-distribution of TSP.

1.2 E-distributions and simulated data

Now consider the uncertainties in the evidence used to obtain the computed trip setpoint $\exp(W)$ using the EVS 2010 method. These are E-uncertainties, characterised by E-distributions. The EVS terminology is 'epistemic' uncertainties and distributions. The equations for applying the E-uncertainties are given on page 19 of the EVS 2010 (updated) document, but are not numbered. Note that in each of these equations, as in equations (10) to (12) for A-uncertainties, the E-errors are applied multiplicatively. For instance, the equation applying error ε^{soro} to the ripples Q has the form $S = Q(1 + \varepsilon^{soro})$.

The technical document sets out some E-distributions for the various error terms arising in the use of the physics codes SORO, RFSP and TUF. However, the NOP benchmark problem used somewhat different distributions for the base case, and then these were further varied in the test suites. The base case E-distributions for the error terms e^{soro}, ε_C^{rfsp}, ε_D^{rfsp}, ε^{ccp} and ε^{dr} are set out in Table 2.

	Specific error SD	Common error SD	RFSP error SD
Error in DM, ε^{dr}	0.00251	0.0	
Error in RFSP simulation of COP, ε_C^{rfsp}	0.01435	0.005	)
Error in RFSP simulation of FOP, ε_D^{rfsp}	0.01655	0.005	) 0.005
E-error in critical channel power, ε^{ccp}	0.02415	0.0085	)
Error in SORO radial flux shape, ε^{soro}	0.01	0.0	

Table 2. Base case standard deviations (SD) of E-distributions for error terms

The first change here from the corresponding table in the technical document was to remove the column for mean error because all of these error terms are assumed to have zero means (and to be distributed normally). The second small change was to combine the three components of error in DM into a single distribution. A more substantial change was to introduce two kinds of common errors in the distributions of ε_C^{rfsp}, ε_D^{rfsp} and ε^{ccp}. These not only acknowledge the fact that there will be common errors associated with COP, FOP and CCP, but also that since the physics code RFSP is used in the computation of all of these terms there will a common error acting across all three vectors.

The following equations clarify the precise meaning and implementation of these errors.

$$\begin{aligned}\varepsilon^{rfsp}_{C\_j} &= \eta^{rfsp}_{C\_j} + \delta^{rfsp}_{C} + \rho \ , & \text{(A)}\\ \varepsilon^{rfsp}_{D\_j} &= \eta^{rfsp}_{D\_j} + \delta^{rfsp}_{D} + \rho \ , & \text{(B)}\\ \varepsilon^{ccp}_{j} &= \eta^{ccp}_{j} + \delta^{ccp} + \rho \ . & \text{(C)}\end{aligned}$$

In these equations, the η terms are the specific errors, the δ terms are the individual common errors, while ρ is the RFSP common error. All terms are independent, normally distributed with zero means and base case standard deviations given in Table 2. Notice that there are separately sampled specific errors for each element of the COP, FOP or CCP vectors, but the common errors δ^{rfsp}_{C}, δ^{rfsp}_{D}, δ^{ccp} and ρ are all single values. The standard deviations for these specific and common errors in Table 2 are such that the overall error standard deviations are the same as in the technical document.

Generating simulated data and computed trip setpoints

The data to which the EVS method is to be applied were generated as follows.

First a set of N true ripples vectors Q_1 to Q_N were obtained by randomly sampling from the 1551 possible values (with replacement). Two ripples sample sizes, $N = 100$ and $N = 500$, were used in the benchmark tests. These sampled true vectors were then disturbed by the (multiplicative) E-errors ε^{soro}_{j} sampled independently according to whatever specific error standard deviation applied for the test case in question. The resulting simulated observations are denoted by S_1 to S_N.

Then for each of the 70 flux shapes the assumed true vectors COP, FOP and CCP^0 were disturbed by the E-errors ε^{rfsp}_{C}, ε^{rfsp}_{D} and ε^{ccp}, respectively. These errors are specified in equations (A), (B) and (C), with each term sampled according to whatever specific and common error standard deviations apply for the test case in question.

Finally, the detector flux values DM were disturbed by applying the E-errors ε^{dr} to dr (which have their default value of 1). These errors were also sampled independently according to whatever specific error standard deviation applied for the test case in question.

The EVS method computes 'best estimate' values $TSP^{0,be}$ for the ideal trip setpoint using equation (18), where C^0 is the disturbed value of CCP^0, S is the disturbed value of Q, COP and FOP denote their disturbed values and DM takes the place of FX in equation (5). EVS compute*s* this for all combinations of the N ripples samples and the 70 flux shapes. (Notice that we did **not** re-perturb the COP, FOP and CCP^0 computations for each sampled ripple.)

The EVS method was applied to calculate the computed trip setpoint $\exp(W)$ from these simulated data.

2 The NOP tests

The NOP tests comprised a base case and three suites of variations. In each test case, two ripples sample sizes were used, $N = 100, 500$.

In each test, $\exp(W)$ was computed a large number of times, each time with a fresh random sample of N ripples vectors and fresh random disturbances ε^{soro}, ε_C^{rfsp}, ε_D^{rfsp}, ε^{ccp} and ε^{dr}.

2.1 Base case

The base case was defined by setting common and specific error standard deviations to values given in Table 2, and using equations (A), (B) and (C). The standard deviations assumed by the EVS method were the same, so that there was no mis-specification.

2.2 Suite 1

Suite 1 comprised 9 tests. Tests 1.1 to 1.6 explored the effect of increasing E-error standard deviations. Tests 1.7 to 1.9 varied the amount of common error, while keeping total error standard deviations unchanged. In all of these there is no mis-specification, so that the assumed standard deviations to be used in the EVS method are the true values. Note, however, that a typographical error in the original specification defined the standard deviation of ρ to be 10 times the originally intended value. The test was run before this error was discovered, and consequently fresh data were obtained under the intended specification, which is now identified as test 1.9*

1.1 The SDs for the common error terms δ_C^{rfsp}, δ_D^{rfsp}, δ^{ccp} and ρ are doubled to 0.01, 0.01, 0.017 and 0.01 respectively.

1.2 The SDs for the specific error terms $\eta_{C\_j}^{rfsp}$, $\eta_{D\_j}^{rfsp}$ and η_j^{ccp} are doubled to 0.0287, 0.0331 and 0.0483 respectively.

1.3 Both of the changes in tests 1.1 and 1.2 are applied, so that the common error and specific error SDs for ε_C^{rfsp}, ε_D^{rfsp} and ε^{ccp} are all doubled.

1.4 The SD of specific errors for ε^{soro} is doubled to 0.02.

1.5 The SD of specific errors for ε^{dr} is doubled to 0.00502.

1.6 The changes in tests 1.3, 1.4 and 1.5 are all applied so that all SDs in Table 2 are doubled.

1.7 The RFSP common error ρ is set to zero (i.e. zero SD), while the SDs for δ_C^{rfsp}, δ_D^{rfsp} and δ^{ccp} are increased to 0.00707, 0.00707 and 0.01 respectively.

1.8 The common errors δ_C^{rfsp}, δ_D^{rfsp}, δ^{ccp} and ρ are all set to zero (i.e. zero SDs), while the SDs for the specific errors $\eta_{C\_j}^{rfsp}$, $\eta_{D\_j}^{rfsp}$ and η_j^{ccp} are increased to 0.016, 0.018 and 0.0261 respectively (as in the technical document).

1.9 The SDs for the common errors δ_C^{rfsp}, δ_D^{rfsp}, δ^{ccp} and ρ are set to 0, 0, 0.015 and 0.1 respectively, while the SDs for the specific errors $\eta_{C\_j}^{rfsp}$, $\eta_{D\_j}^{rfsp}$ and η_j^{ccp} are decreased to 0.0125, 0.015 and 0.0189 respectively.

1.9* The SDs for the common errors δ_C^{rfsp}, δ_D^{rfsp}, δ^{ccp} and ρ are set to 0, 0, 0.015 and 0.01 respectively, while the SDs for the specific errors $\eta_{C\_j}^{rfsp}$, $\eta_{D\_j}^{rfsp}$ and η_j^{ccp} are decreased to 0.0125, 0.015 and 0.0189 respectively.

2.3 Suite 2

This suite comprised 7 tests looking at mis-specification of the magnitudes of E-error terms, with assumed error standard deviations used by the EVS method differing from the true values that are used to generate the simulated data. In each test the true values are as in Table 2. Test 2.6 was originally run with the erroneous specification of test 1.9, and fresh results were obtained using the intended specification, which is denoted as test 2.6*.

2.1 Assume all SDs to be 25% larger than in Table 2.

2.2 Assume all SDs to be 20% smaller than in Table 2.

2.3 Assume all common error SDs to be 25% larger than in Table 2.

2.4 Assume all common error SDs to be 20% smaller than in Table 2.

2.5 Assume error SDs as in test 1.8.

2.6 Assume error SDs as in test 1.9.

2.6* Assume error SDs as in test 1.9*.

2.4 Suite 3

This suite comprised 8 tests looking at changing the weights applied to the 70 flux shapes.

3.1 All 70 flux shapes have weights 0.0142857.

3.2 The 14 flux shapes with base case weights 0.0274725 have their weights reduced to 0.0214286, while the remaining 56 have their weights increased to 0.0125.

3.3 The 14 flux shapes with base case weights 0.0274725 have their weights reduced to 0.013179, flux shapes 21, 22, 23, 24, 40, 56, 57, 58, 59 and 75 have their weights increased to 0.031, while the remaining 46 have weights unchanged at 0.010989.

3.4 The 20 flux shapes 9, 13, 17, 21, 25, 29, 33, 37, 40, 41, 44, 48, 52, 56, 60, 64, 68, 72, 75 and 76 have weights 0.05, while the remaining 50 have zero weights.

In the above 4 tests, as in the base case, no distinction is made between the true weights which form the A-distribution of φ and the assumed weights which are to be used in the EVS method. There is therefore no mis-specification. Tests 3.5 to 3.8 are defined by setting the A-distribution of φ as in tests 3.1 to 3.4 but with the assumed weights used in EVS set to the base case values (and hence implying mis-specified weights).

3 Performance measures

Equivalent performance measures were used in this benchmarking exercise to those employed in the Benchmark A tests.

- *Mean.* The mean of the M solutions.
- *SD.* The standard deviation of the M solutions.
- *Non-coverage.* The proportion of the M solutions that lie below $t_{.95}$. Ideally, this should be 5%.
- *Mean deficit.* For every $W^{[m]}$ that is above $t_{.95}$, the deficit is the proportion of $CP^{(d)}_{\max}$ values that lie below $W^{[m]}$. The mean deficit is the average of the deficit values for all $W^{[m]}$ above $t_{.95}$. By convention if no values are above $t_{.95}$ then the mean deficit is 0.05.

4 Performance criteria

The error in the specification of test 1.9 implied extreme values of errors in the RFSP-based computations, and of correlations between those errors. Tests 1.9 and 2.6 are anomalous in the sense that we would not expect such levels of error in practice and so do not require EVS 2010 to meet the performance criteria on these tests.

Performance of EVS 2010 against the six performance criteria is assessed in the next two subsections, ignoring the anomalous tests 1.9 and 2.6. Results from those tests are discussed briefly in the final subsection.

4.1 Tolerance limit criteria

There are three criteria in this group.

- *Criterion 1 (Desirable).* The non-coverage should be 5% or less in the base case and all tests where there is no mis-specification. It should also be less than, or not much more than, 5% in tests where the mis-specification is minor.

Tests with no mis-specification include all of Suite 1 and tests 3.1 to 3.4 in Suite 3, including test 1.9* but ignoring test 1.9. The tests with mis-specification are all of Suite 2 and tests 3.5 to 3.8 in Suite 3, including test 2.6* but ignoring test 2.6. Non-coverage is well below 5% for all tests and both values of N. EVS 2010 passes Criterion 1.

- *Criterion 2 (Essential).* In all tests to which Criterion 1 applies, non-coverage must not be excessive *and* when non-coverage exceeds the levels set out in Criterion 1 the mean deficit must not be excessive.

EVS 2010 passes Criterion 1 and so automatically passes Criterion 2.

- *Criterion 3 (Desirable).* In the base case and all tests where there is no mis-specification, the non-coverage should be closer to 5% than to 0%.

EVS 2010 clearly fails Criterion 3, since non-coverage is zero or close to zero for all relevant tests. Criterion 3 is not essential and this performance is acceptable although it indicates inefficient use of available information.

4.2 Face validity criteria

This group also comprises three criteria.

- *Criterion 4 (Essential).* As the sample size N increases in any test, the mean must increase and the SD must decrease.

The required increases/decreases are observed in all tests. EVS 2010 passes Criterion 4.

- *Criterion 5 (Desirable).* If E-error variances increase, then the mean should decrease and the SD should increase.

Suite 1 provides a variety of comparisons to which this criterion applies. All of the tests 1.1 to 1.6 have higher variances than the base case, test 1.3 has higher values than in tests 1.1 and 1.2, and test 1.6 has higher variances than all of 1.1 to 1.5. On a total of 9 such comparisons, we observe that in many cases the required movements clearly occur. The very small reductions in variance in tests 1.4 and 1.5 relative to the base case may simply be due to sampling variation.

However, there is an interesting pattern when we consider the change from the base case to test 1.1 and from test 1.2 to 1.3. These two comparisons are associated with increasing the common errors in RFSP-related computations. In both comparisons and for both values of N we see small increases in the mean output, contrary to the requirement of this criterion. Although the changes are small there is a pattern here and in one case at least the change is not explainable by chance variation. This is the change of mean from 123.7 (base) to 124.0 (test 1.1) for $N = 500$, which is 4 standard deviations away from zero (for 1000 iterations).

EVS 2010 fails Criterion 5. However, this can be discounted if the next criterion is met.

- *Criterion 6 (Essential).* If Criterion 5 is not met for any comparison of tests, then it must be clear that for all reasonable values of the E-error variances the non-coverage (and mean deficit) will remain acceptable.

Although the behaviour of EVS 2010 when common RFSP errors are increased is paradoxical and fails Criterion 5, the changes are nevertheless small and in my judgement it is unlikely that for plausible values of error variances in practice would lead to non-coverage appreciably exceeding 5%. EVS 2010 passes Criterion 6. However, see the discussion of varying correlations in section 5.1.

4.3 The anomalous tests

Non-coverage exceeds 50% in test 1.9 for both values of N and in test 2.6 it exceeds 95% for both values of N. Comparing test 1.9 with test 1.1, the increased variance should result (Criterion 5) in a decrease in the mean, whereas the mean is actually much higher. Indeed it is markedly higher (131.1 at $N = 100$ and 130.6 at $N = 500$) than the true reference value of $t = 127.8$. (The non-coverage is only a little over 50% because the variance is also high.) This would constitute a very strong violation of Criterion 5 and a failure of Criterion 6 if this test had been included in the main analysis. It is worth remembering that there is no mis-specification in test 1.9; the variance of ρ is high but its value is correctly specified. In test 2.6, the variance is wrongly specified to be high when in fact it is low. It has been noted before that EVS 2010's performance can be poor when error variances are over-estimated, and this is an extreme case. In practice, variances of physics code errors are estimated from validation data. The impact of this finding is that instead of using standard (e.g. unbiased) estimates as the assumed values in EVS, the estimation errors should be quantified (allowing for the fact that validation is, as I understand it, typically of code predictions against observations which themselves are subject to observational errors) and estimates reduced accordingly.

It is clear that if extreme (and for the NOP problem, perhaps quite unrealistic) levels of E-error exist then EVS 2010's performance will be unacceptable, even when there is no mis-specification.

5 Other analyses

Some of the tests in this exercise were introduced to study other forms of robustness.

5.1 Varying correlation

The assumed E-error models for the computation codes that have apparently been used in previous demonstrations of EVS 2010 for the NOP problem have apparently been simpler than that specified for the base case in these tests. In particular, the technical document given to me specified zero variances for the common errors δ_C^{rfsp}, δ_D^{rfsp}, δ^{ccp} and ρ. As a result, the computation codes were treated as producing independent errors, even for adjacent channels. Although I understand that this error structure has been the basis of some previous computations, I took the view that it was unrealistic. I introduced non-zero values for all of these variances in the base case, thereby creating correlations. Although I believe that correlations may have been used in some other previous analyses by AMEC NSS, I think that my overall common error ρ had not featured before. It is clear that the nature and magnitude of correlations in E-errors is a topic that is still poorly understood and therefore in practice would be prone to mis-specification. It was with this in mind that I included tests which kept the overall magnitude of error constant while varying the degree of correlation. Test 1.7 removes the overall common error term ρ, while test 1.8 removes common error entirely (the case originally specified in the technical document). Test 1.9* goes the other way, increasing the amount of correlation. (The erroneously specified test 1.9 actually has even stronger correlation, but its usefulness is compromised by the fact that overall error variance is also increased markedly.)

The performance of EVS 2010 on tests 1.7, 1.8 and 1.9* is again somewhat paradoxical. As the degree of correlation is increased, from test 1.8 to 1.7 to the base case to 1.9*, the mean output increases and the SD also increases. This kind of change is worrying. It gives cause for concern that if correlation were to be increased further relative to test 1.9* we might see the output mean and SD both increasing to a level that could lead to unacceptable performance. In test 1.9* the effect is already to produce a non-zero non-coverage value. As has been remarked upon previously, EVS 2010 tends to under-estimate the reference TSP and this leads to non-coverage being estimated as zero in almost all tests. The benchmarking has suggested that EVS 2010 performance may be sensitive to the specification of correlations, with the possibility that sufficient correlation might lead to unacceptably high non-coverage and/or mean deficit.

Notice that the above discussion is all about correctly specified analyses. In contrast, tests 2.5 and 2.6* were intended to examine the effect of mis-specifying the degree of correlation. Test 2.5 is the more interesting case because it assumes no correlation when correlation is really present. Although the outcome of this test is acceptable in the sense that non-coverage is zero, the figures do indicate some cause for concern. If we compare the mean and SD for tests 1.8 and 2.5 we see that in 2.5 both measures are higher (for both values of N). Although the changes are smaller, there is a suggestion of the same effect moving from the base case to test 2.5. Test 2.6* confirms this trend. In general, increasing both mean and SD is behaviour that has been noted as paradoxical and potentially worrying in several other contexts. In general, it seems that EVS 2010 will be sensitive to mis-specification of correlation structure. This is a matter of concern in practice because it seems to me that correlations have not been seriously considered hitherto.

It is not possible to say whether correlations within the range of realistic values might lead to unacceptable behaviour because the whole correlation structure of the problem is extremely complex. In particular, it is likely that E-errors in physics code computation are spatially correlated, a possibility which has not been explored at all in these tests.

5.2 Specification of flux shapes

Test suite 3 was designed to study the effect of changes in the flux shape distribution. Four alternative sets of probabilities (or weights) for the 70 flux shapes in the NOP benchmark problem were rather arbitrarily chosen. In tests 3.1 to 3.4 these alternative distributions apply and are correctly specified. In tests 3.5 to 3.8 the base case weights were assumed but the true weights were as in the alternative distributions, so these tests involved mis-specification. In all cases, non-coverage remained firmly below 5% and there were no paradoxical movements of mean and SD. EVS 2010 apparently performs well under variations and mis-specifications of the flux shape weights, within the range explored in these tests.

6 Conclusions

The NOP benchmark scenario is about as close to a real NOP trip setpoint application as it is feasible to create for benchmarking purposes, yet it is not a perfect representation. The assumed true values of ripples and flux shapes have really been obtained from physics code computations and as such are likely to be more variable and less smooth that true values. Also the assumed correlation structure, whilst already more complex than has apparently been used before, is probably still not as realistic as should be employed in real applications. To the extent that the NOP benchmark exercise is a good measure of how EVS will perform in real NOP trip setpoint applications, this report draws the following conclusions.

1. EVS 2010 passes all the essential performance criteria for tolerance limit validity and face validity in the context of changes in E-error variances.
2. EVS 2010 performs well and robustly in the context of changes in flux shape distribution.
3. Further instances of so-called paradoxical behaviour continue to arise, as they have in the previous benchmarking tests. These can generically be

described as instances where a change in specification leads to an increase in both mean and SD. Such changes always flag a warning because if increases were large enough they would almost inevitably lead to poor tolerance limit performance (excessive non-coverage and/or excessive mean deficiency). They have been found under conditions of correct specification as readily as under mis-specification. None of these instances to date has led to poor tolerance limit performance in the actual tests, but of course the tests do not explore all possibilities.

Criterion 6 asks whether excessive non-coverage or mean deficit might arise under real conditions, and to the extent that I am able to judge this I think EVS 2010 has enough built-in robustness (through its tendency to underestimate TSP) to retain acceptable performance very widely. However, I am not really qualified to judge this, and the possibilities in terms of error correlations remain to be examined properly.

Appendix — Test results

RSP-0293

Table10_conv.txt									
Test	**t(0.95)**	**mean(TL)**		**std(TL)**		**non-cov**		**mean-def**	
		100	500	100	500	100	500	100	500
Base	127.81	122.5	123.7	1.27	1	0	0	5	5
1.1	127.81	122.7	124	1.91	1.74	0.3	1.7	7	8
1.2	127.81	118.8	120	1.94	1.65	0	0	5	5
1.3	127.81	118.9	120.1	2.56	2.32	0	0.1	5	8.5
1.4	127.81	121.7	123	1.26	0.99	0	0	5	5
1.5	127.81	122.4	123.7	1.27	0.99	0	0	5	5
1.6	127.81	118.4	119.5	2.51	2.22	0	0	5	5
1.7	127.81	122.5	123.6	1.16	0.83	0	0	5	5
1.8	127.81	122.3	123.4	1.02	0.72	0	0	5	5
1.9	127.81	131.1	130.6	13.54	13.66	58.3	57.6	68.4	70.7

1.9*	127.81	123.1	124.4	1.77	1.56	0.3	1.2	6.3	8

* Test 1.9* uses the intended specification for test 1.9

RSP-0293

Table20_conv

Test	t(0.95)	mean(TL)		std(TL)		non-cov		mean-def	
		100	500	100	500	100	500	100	500
Base	127.81	122.5	123.7	1.27	1	0	0	5	5
2.1	127.81	124.7	125.8	1.33	1.02	0.8	2.3	6.4	7.2
2.2	127.81	121	122.3	1.25	0.95	0	0	5	5
2.3	127.81	122.8	123.9	1.25	1.02	0	0	5	5
2.4	127.81	122.4	123.6	1.28	0.99	0	0	5	5
2.5	127.81	122.6	123.9	1.27	1.01	0	0	5	5
2.6	127.81	129.3	129.8	0.85	0.81	95.4	99.1	15.2	19.7
2.6*	127.81	120.2	121.3	1.23	0.99	0	0	5	5

* Test 2.6* uses the intended specification for test 2.6

Table30_conv

Test	t(0.95)	mean(TL)		std(TL)		non-cov		mean-def	
		100	500	100	500	100	500	100	500
3.1	127.83	122.7	123.8	1.18	0.95	0	0	5	5
3.2	127.82	122.6	123.7	1.23	0.96	0	0	5	5
3.3	127.8	122.9	124	1.25	0.97	0	0	5	5
3.4	127.63	122.2	123.3	1.6	1.39	0	0.2	5	5.4

RSP-0293

Table4

Test	**t(0.95)**	**mean(TL)**		**std(TL)**		**non-cov**		**mean-def**	
		100	500	100	500	100	500	100	500
3.5	127.83	122.5	123.7	1.27	1	0	0	5	5
3.6	127.82	122.5	123.7	1.27	1	0	0	5	5
3.7	127.8	122.5	123.7	1.27	1	0	0	5	5
3.8	127.63	122.5	123.7	1.27	1	0	0	5	5

RSP-0293

OPG/BP/AMEC NSS Comments and AO'H Dispositions

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO'H dispositions** *+ Actions to revise report*
Final Report, Page 3 Executive Summary Page 33 Summary and Conclusions	*If there is concern over the assumption of superposition and independence for this application, robustness of the EVS2010 method to plausible departure from it should be investigated (or a version of EVS developed which does not rely on the assumption).*	• The superposition and independence considerations have not been evaluated as part of the scope of the research project. Since the issue has not been discussed and established as a concern, recommend this statement be removed from the "Executive Summary" and "Summary and Conclusions" sections of the report.	The scope of the contract included assessment of fitness for purpose. This required all assumptions and approximations made in the theory and application of EVS to be examined in the benchmarking. It is legitimate for the final report to draw attention to issues which are unresolved in the benchmarking and which may in some circumstances give rise to concern. *Action: None required.*
Final Report, Page 32, Summary and Conclusions	*If there is concern over the assumed A-distribution of θ, over the weights in the A-distribution of Φ, or over the assumed forms of the E-distribution of physics code errors, robustness to their mis-specification should be tested.*	• The issue of aleatory distributions and their independence have not been evaluated as part of the scope of the research project. Since the issue has not been discussed and established as a concern, recommend the reference to concern to the A-distributions be removed from the "Summary and Conclusions" section of the report	The scope of the contract included assessment of fitness for purpose. This required all assumptions and approximations made in the theory and application of EVS to be examined in the benchmarking. It is legitimate for the final report to draw attention to issues which are unresolved in the benchmarking and which may in some circumstances give rise to concern. *Action: None required.*

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO'H dispositions** *+ Actions to revise report*
Final Report, Page 32 Summary and Conclusions Finding #3	*The EVS 2010 theory is a rather complex development of a tolerance limit. This complexity makes it difficult to understand the mechanisms by which certain kinds of behaviour arise. A somewhat simpler solution would be preferable.*	• The NOP problem is a complex problem and it is unlikely that a solution would look any less complex. • The existing tolerance limit solutions (i.e., parametric or non-parametric) appear to be simpler, however, they are more suitable to solving simpler problems. For those problems, the only source of uncertainties is the finite sample size. • EVS method is a unique tolerance limit solution which in addition to the finite sample size, also treats epistemic error separately from aleatory error. • Simpler methods (e.g., order statistics and the traditional methods) do not give sufficiently accurate results.	The intention here was to refer to the way that EVS constructs a tolerance limit from the 'plug-in' or 'best estimate' TSP (U or V in the theory). In my view this is an unnecessarily indirect approach to using the data. However, I have no serious alternative in mind and, to be fair, other approaches would probably introduce other complexities. *Action: Remove "A somewhat simpler solution would be preferable."*
Final Report, Finding #4	*Like any statistical method, the EVS 2010 theory relies on a number of assumptions which might not hold in practice. In particular the assumption of superposition and independence is acknowledged to be at best an approximation.*	• We agree that these are important issues. However, these are general issues independent of the statistical method (e.g., TM, BEAU, EVS, order statistics, etc.). EVS is not dependent on the assumptions associated the superposition/independence. • The superposition considerations have not been evaluated as part of the scope of the research project. Since the issue has not been discussed and established as a concern, recommend this statement be removed from the final report. • The issue of aleatory distributions and their independence have not been evaluated as part of the scope of the research project. Since the issue has not been discussed and established as a concern, recommend the reference to concern to the A-distributions be removed from the final report.	The scope of the contract included assessment of fitness for purpose. This required all assumptions and approximations made in the theory and application of EVS to be examined in the benchmarking. It is legitimate for the final report to draw attention to issues which are unresolved in the benchmarking and which may in some circumstances give rise to concern. *Action: None required.*

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO'H dispositions** *+ Actions to revise report*
Final Report, Finding #5	*The EVS 2010 theory assumes that the structures of error distributions are known. There is potentially great complexity in the variance and covariance structures for the error vectors, whereas in practice rather simple structures (equal variances, equal or zero correlations) have been assumed. The benchmarking suggests that the details of these structures may be important, but little is known about the behaviour of EVS 2010 under variations and mis-specifications of these structures.*	The error structure assumed for the benchmarking exercise is a simplified representation for the purpose of understanding the EVS method. These are important points raised which we have investigated (e.g., Darlington NOP Uncertainty Analysis Report (Reference [**Error! Reference source not found.**]) documented the identification, classification, quantification and justification of errors). • NOP benchmarking exercises showed that the impact on the Tolerance Limit due to changes in error structure is small, even with mis-specification: Example: All suite 2 cases (except 2.6) had low non-coverage and small mean deficits for all sample sizes. • Results for MCP Group B exercises were shown to be sensitive to mis-specification but since then been addressed (see Reference [**Error! Reference source not found.**]). • EVS does not impose any restrictions on the error structure. • There is nothing inherent in EVS that requires simple error structures such as equal variances, equal or zero correlations. In practice, the error structures for all input variables are derived using qualified validation datasets and estimation methods which are consistent with statistical theory. • EVS results behave/respond naturally to different complexities in the error structure of each input.	• The change in the Mean in the NOP benchmark tests from the base case to test 2.6* is more than two SDs. This is not a small change, and if EVS were not so 'biased' it would have a very marked effect on Non-coverage. This is the basis of the claim that the details of the correlation structure may be important. • It may be that more complex structures have been explored in a report that I have not seen, but these have not been evaluated in the benchmarking *Action: None required.*
Final Report, Finding #6	*The EVS 2010 theory also assumes that the values of various constants are known. In practice these will not be known*	• It is a valid concern and efforts to address this issue has been completed in addressing ITP review comments (e.g., Cory Atwood). To implement the theory numerically, approximations to the parameters used in	I accept that some extensive testing of the effect of the surrogate method has been carried out and reported in

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO'H dispositions** *+ Actions to revise report*
	and can at best be estimated. A so-called surrogate methodology is employed for some of these, which involves several levels of approximation, and therefore the quality of estimates may be poor. The EVS 2011 Report introduces a calculation of the variance of a key part of the trip setpoint computation arising from estimation errors, but this is then used only to make an ad hoc adjustment which does not properly account for the uncertainty introduced by estimation.	the theory are made based on available data. Extensive testing is performed and great care is taken to ensure that our estimates are adequate. • Additional issues raised in the report have now been investigated and confirmed the adequacy of the existing estimates with EVS-2010. It can be shown that the variance in $\Delta\lambda$ can be used directly in the calculation (i.e., there is no need for ad hoc assumptions). This issue has been addressed (see Reference [**Error! Reference source not found.**]).	the EVS 2010 document. I do not know the basis of the claim in the second bullet point. The ad hoc assumption referred to is to apply an adjustment by a multiple of the standard deviation of λ. Although this is apparently modified in a version of the EVS methodology that has not been evaluated, it is an intrinsic part of the EVS 2010/2011 methodology that is the subject of this contract. *Action: Remove "and therefore the quality of estimates may be poor."*
Final Report, Finding #7	*In practice, EVS 2010 appears to be very sensitive to mis-specification of the standard deviations of computational errors in physics codes. Even a modest over-estimation of these standard deviations can lead to unacceptable levels of non-coverage.*	• NOP benchmarking exercises showed that the impact on the Tolerance Limit due to changes in error structure is small, even with mis-specification (e.g., All suite 2 cases (except 2.6) had low non-coverage and small mean deficits for all sample sizes). The issue of mis-specification is still an important point which we have investigated (e.g., Darlington NOP Uncertainty Analysis Report (Reference [**Error! Reference source not found.**]) documented the identification, classification, quantification and justification of errors). Specifically, Reference [**Error! Reference source not found.**] provides justification of the	The change in the Mean in the NOP benchmark tests from the base case to test 2.1 is about two SDs. This is not a small change, and if EVS were not so 'biased' it would have a very marked effect on Non-coverage. Even with the 'bias' we see in 2.1 the highest Non-coverage of any test in the NOP suites (except 1.9 and 2.6).

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO'H dispositions** *+ Actions to revise report*
		error specification and conservative values are adopted. • Results from MCP Group B exercises were shown to be sensitive to mis-specification but since then been addressed (see Reference [**Error! Reference source not found.**]). The investigation carried out in Reference [**Error! Reference source not found.**] produce results which demonstrate that EVS is not overly sensitive to mis-specification.	The reference in the second bullet point again appears to be to a revised methodology that has not been evaluated. *Action: None required.*
Final Report, Finding #8	*The benchmarking exercise has revealed a variety of instances of what has been called paradoxical behaviour in some of the appendices to this report but is here referred to as failures to meet the Mean and SD Consistency Criterion. All such instances raise the potential for excessive non-coverage. The positive finding 2 holds because excessive non-coverage has not generally been found in the benchmarking tests, but the potential remains if parameters are set in practice outside the range explored in the benchmarking exercises. The fact that one such instance arises when varying correlations in E-distributions of physics code*	• Current EVS -2010 provides adequate estimates (i.e., conservative NOP TSP results) as confirmed in NOP benchmarking tests i.e., passes all Essential performance criteria #2, #4, and #6. • For the MCP problem, the changes in the mean tolerance limit is small. The results from suite 1.7 to 1.3 produced mean tolerance limits of 6871 kW and 6870 kW, respectively. Note that the true value was 6817 kW. • For the NOP problem, the changes in the mean tolerance limit is also small. The results from (suite 1.7 to 1.1) produced mean tolerance limits of 123.6 %FP and 124.0 %FP, respectively. Note that the true value was 127.8 %FP. • Efforts to further investigate this review comment was completed. The recent investigations show: o the mean tolerance limit behaves in accordance with the stated expectation as a function of epistemic uncertainty. Therefore, the observed trend in review finding #8 is no	I accept that observed instances of failure of the Mean and SD Consistency Criterion typically involve small changes in Mean. This has never been disputed. Nevertheless, in the words that I used when presenting my report on 25 February 2013, when these face validity criteria fail all bets are off. It is no longer possible to anticipate the behaviour of the method when applied outside the range of parameters which have been studied in the benchmarking. That range itself is necessarily very limited. This is the basis for

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO'H dispositions** *+ Actions to revise report*
	errors enhances the importance of negative finding 5.	longer applicable. ○ there is no impact on existing results: NOP TSPs remain conservative and passes all essential performance criteria. • This issue has been addressed and described further in Reference [**Error! Reference source not found.**].	my proposal of "empirical support studies" (which was made informally in the presentation and will be included by request of CNSC in the revised final report). *Action: None required.*
Final Report, Finding #9	*EVS 2010 in practice seems to have a 'bias', specifically a tendency to produce computed TSP values that are unnecessarily far from the reference value and so results in non-coverage values that are in almost all benchmarking tests well below the nominal 5%. This has the beneficial effect of providing a cushion to prevent non-coverage becoming excessive under conditions of mis-specification or paradoxical behaviour, which is shown in the positive finding 2. However, there are some indications that the 'bias', and therefore also its beneficial effect, may reduce as the size of the sample of ripples "observations" increases. The 'bias' also means that EVS 2010 is processing the available information rather inefficiently.*	• Conservative results are desirable in safety related problems. • Indeed EVS non-coverage results are below the 5%. However, this points to a conservative estimation which is required. Nevertheless, the obtained results are quite close to the true values. • Efforts to further investigate this review comment was completed (Reference [**Error! Reference source not found.**]). The recent investigations show: ○ no reduction in the EVS performance as the sample size increases. ○ EVS NOP TSP results remain conservative (low non-coverage and small mean deficit).	As already mentioned, Reference [1] appears to concern a modification of EVS methodology which has not been evaluated. *Action: None required.*

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO'H dispositions** *+ Actions to revise report*
MCP Report, Principle Finding #1	*In the MCP benchmark tests, EVS performance in regard to the tolerance limit Criteria 1 and 2 is found to be poor for low sample sizes. N = 100 does not appear to be large enough to satisfy these criteria consistently, even when the error variances are correctly specified.*	• For practical applications, N is much larger than 100 (e.g. 500 or more ripples). • Efforts to further investigate this review comment was completed (Reference [**Error! Reference source not found.**]). The recent investigations show: o Non-coverage probabilities for all Suite 1 cases and for all sample sizes (including low sample sizes such as N = 20 or N = 100) are below 5%. Furthermore, the mean deficit is shown to improve as well and is satisfactory in all cases. o These conclusions apply to Suite 2 exercises (except for 2.4 and 2.8 for which the mis-specification is excessive).	The finding is pertinent and stands because previous tests had suggested that N = 20 was too small but N = 100 produced acceptable performance. I accept that in practice N may be expected to be well above 100, but this does not change the finding. *Action: None required.*
MCP Report, Principle Finding #2	*In the MCP benchmark tests, EVS performance on the tolerance limit Criteria 1 and 2 is highly sensitive to mis-specification of error variances. It fails Criteria 1 and 2 when error variances are overstated by as little as a factor of 1.25. Performance becomes worse with increasing sample size. This is a serious concern, because in practice it will not be easy to ensure that error variances are not over-estimated.*	• Note that for the NOP problem, the mean tolerance limits are not overly sensitive to mis-specification e.g., In NOP Benchmarking: All suite 2 cases (except 2.6) had low non-coverage and small mean deficits for all sample sizes. • However, care was taken to the specification of each error estimates for all NOP analyses. Uncertainty analysis report documents the quantification and justification of each error estimates. • Although the current EVS method for NOP is adequate, additional efforts to further investigate this review comment was completed (Reference [**Error! Reference source not found.**]). • To address the finding noted in the MCP problem: the recent investigations show that the non-coverage probabilities and mean deficit are no longer highly sensitive to mis-specification. In particular: o Suites 2.3 and 2.7 results were of great concern since their mis-specification of the error by a	The first bullet point has already been answered. Also, Reference [1] appears to concern a modification of the methodology which has not been evaluated. *Action: None required.*

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO’H dispositions** *+ Actions to revise report*
		factor of 1.25 is not very large, yet the corresponding non-coverage and mean deficit were excessively large. It was shown that this is no longer true. The non-coverage in both cases is under 5%. The corresponding mean deficit is satisfactory as well. ○ The concern regarding the performance with increasing sample size is also no longer an issue. It is shown that the non-coverage and mean deficit are consistently satisfactory for all sample sizes. For this reason, we included a sample size of N = 1000 and showed a good performance for such a large sample size as well.	
MCP Report, Principle Finding #3	*In the MCP benchmark tests, EVS exhibits paradoxical behaviour which fails Criterion 5. Although I deem that this behaviour is not sufficiently marked for EVS to fail the essential Criterion 6, there remain two causes for concern. First, the emergence of paradoxical movements of mean values when error variances change (without being mis-specified) appears to be a widespread and intrinsic problem with EVS. Second, although these effects have generally been found to be small, there is evidence that they become more marked with increasing sample*	• Efforts to further investigate this review comment was completed (Reference [**Error! Reference source not found.**]). The recent investigations show that the mean tolerance limit behaves in accordance with the stated intuition as described in Criterion 5. ○ For all Suite 1 cases and for all sample sizes, the mean tolerance limit increases if either the variance of the channel common error increases or the variance of the channel specific errors increases. ○ A larger sample was included (i.e., N = 1000) to demonstrate that the raised issue regarding the movement of the mean values having a more marked effect with increasing sample size is no longer a concern.	Reference [1] appears to concern a modification of the methodology which has not been evaluated. *Action: None required.*

Reference	**Review Finding**	**OPG, Bruce Power and AMEC NSS Comments**	**AO'H dispositions** *+ Actions to revise report*
	size.		
NOP Report, Conclusion #3	Further instances of so-called paradoxical behaviour continue to arise, as they have in the previous benchmarking tests. These can generically be described as instances where a change in specification leads to an increase in both mean and SD. Such changes always flag a warning because if increases were large enough they would almost inevitably lead to poor tolerance limit performance (excessive non-coverage and/or excessive mean deficiency). They have been found under conditions of correct specification as readily as under mis-specification. None of these instances to date has led to poor tolerance limit performance in the actual tests, but of course the tests do not explore all possibilities.	• Efforts to further investigate this review comment was completed (Reference [**Error! Reference source not found.**]). The recent investigations show that the mean tolerance limit behaves in accordance with the stated intuition as described in Criterion 5: ○ for both common and specific errors the behaviour of the mean tolerance limit decreases. These results hold for all Suite 1 cases and for all sample sizes (i.e., no reduction in the EVS performance as the sample size increases). ○ The NOP TSP results remain conservative (low non-coverage and small mean deficit). ○ The concern of the "so-called paradoxical behaviour continue to arise" is no longer applicable.	Reference [1] appears to concern a modification of the methodology which has not been evaluated. *Action: None required.*

Prepared by Tony O'Hagan

13 March 2013