Overview of the Development of Mitigating Systems Performance Indices (MSPIs)

Draft
April 28, 2003

Table of Contents

List of Tables

List of Acronyms

AOT	allowed outage time
B	Birnbaum
CCF	common cause failure
CDF	core damage frequency
CNIP	constrained non-informative prior
FV	Fussell-Vesely
MLE	Maximum Likelihood Estimate
MSPI	Mitigating System Performance Index
NEI	Nuclear Energy Institute
NRC	Nuclear Regulatory Commission
PI	performance indicator
PRA	probabilistic risk assessment
PSA	probabilistic safety assessment
PWR	pressurized water reactor
RBPI	risk-based performance indicator
RCP	reactor coolant pump
ROP	Reactor Oversight Process
SDP	Significance Determination Process
SPAR	Standardized Plant Analysis – Risk
UA	UnAvailability
UR	UnReliability

1. Purpose

The purpose of this paper is to discuss the technical basis for development of "mitigating systems performance indices" (MSPIs). The discussion will cover the following:

- formulation of MSPIs;
- approximations on which the MSPIs are based;
- benefits of MSPIs;
- limitations of MSPIs, including scope limitations;
- key issues in MSPI development to be addressed in near-term work.

2. Background

The Reactor Oversight Process (ROP) currently uses performance indicators that quantify unavailability. There are certain issues associated with these indicators, including (a) the use of generic thresholds, and (b) the way in which fault exposure time associated with failure events affects the values of the current indicators.

Phase 1 of the Risk-Based Performance Indicator (RBPI) Development program (Ref. 1) explored several possible enhancements to the ROP performance indicators. A key aspect of the Ref. 1 approach was the use of plant-specific models (the SPAR models) to assess the risk significance of changes in unreliability (UR) and unavailability (UA). Based on these models, it was possible to develop candidate RBPIs that separately quantify UR and UA within a common model framework. It was also possible to determine plant-specific thresholds for these indicators. These enhancements help to address the issues mentioned above for current ROP indicators. In the Phase 1 RBPI effort, these enhancements were shown to be generally feasible, although for some UR indicators, statistical uncertainty is an issue.

Although these candidate indicators have certain benefits compared to the performance indicators (PIs) currently in use, they also have certain drawbacks. In particular, implementing separate train-level UR and UA indicators leads to a substantial increase in the number of indicators. This increase in the number of indicators raises concerns regarding increased burden associated with reporting data to support the indicators, and the effect of a larger number of indicators on the action matrix. In addition, including a larger number of indicators increases the likelihood that at least one indicator will give a false indication.

The MSPIs are intended to reap the benefits of the improved treatment developed in the RBPI program (improved quantification of UR, using plant-specific thresholds) while resolving the issues associated with proliferation of indicators. The MSPI approach separately quantifies the significance of changes in UR and UA, but then rolls up these contributions into a single system-level indicator. The MSPI approach does this using a simplified calculational approach based on importance measures, thereby avoiding the need for ongoing manipulations of the entire risk model. This approach is quantitatively adequate until changes in UR and UA become very large, at which point the numerical inaccuracy does not matter, because licensee and regulatory attention has already become focused on these contributions.

At the end of the Phase 1 RBPI development, approaches to higher-level indicators (such as MSPIs) were discussed with ACRS and with external stakeholders, and mentioned in Section 6.5 of NUREG-1753. Since then, MSPIs have been publicly discussed on numerous occasions

(public workshops, etc.), and work has been performed both by NRC and by industry to better understand the pros and cons of MSPIs.

3. Characteristics of MSPI's

3.1 Purpose of MSPI's

The purpose of MSPI's (NEI 99-02, Ref. 2) is to "monitor the performance of selected systems based on their ability to perform risk-significant functions... ." The emphasis on "risk-significant" is aimed at the distinction between the functionality associated with design basis requirements or other prescriptive requirements, and the functionality associated with the mission success requirements in a PRA model.

A good example of what is meant by the "risk-significant function" is that of the Chemical and Volume Control System in a PWR. The normal function of the charging pumps is to provide normal flow during plant operation to control water level and water chemistry (e.g., the boric acid concentration in the reactor coolant system for reactivity purposes). But in many PRAs, the risk-significant function is to provide emergency core cooling in the event of a loss-of-coolant accident, and/or reactor coolant pump (RCP) seal injection flow.

3.2 Preliminary Observations Regarding Importance Measures

The MSPI calculations are much easier to understand, given certain observations regarding "importance measures." This subsection provides those observations.

Suppose that we are interested in element *A*, and suppose that *A* appears in the *CDF* expression multiplying X and multiplying Y.* (*A* could be a basic event corresponding to unreliability or unavailability.) In addition to these *A*-related contributions, there are other contributions *Z* that do not contain *A*. Then we can write

$$CDF = A * X + A * Y + Z$$

The "Fussell-Vesely" "importance" FV of element A can be approximated** as

$$FV(A) = \frac{A * X + A * Y}{CDF}$$

This is the fractional contribution to *CDF* of terms containing *A*. The "Birnbaum" "importance" *B* of element *A* can be written

$$B(A) \equiv CDF(A=1) - CDF(A=0) = X + Y = FV(A) * \frac{CDF}{A}$$

Note that although the value of *A* appears in the denominator of the rightmost expression, the value of *B(A)* is independent of the value of *A*, as shown in the preceding equality. For some

* In this discussion, symbols such as "*A*" are being used to denote either a Boolean event or its probability, depending on context.

** The formulae in this section reflect the "rare-event" approximation. For present purposes, the effort needed to work around this approximation is not justified by the value added.

purposes, it is useful to think of *B(A)* as the partial derivative of *CDF* with respect to *A*. *B(A)* is also the "coefficient" of *A* in the *CDF* expression.

Another noteworthy point to make about *B* is that elements that are logically in series have the same *B* (apart from considerations related to the caveats summarized at the end of this section), although they do not have the same *FV*. The reason that they have the same *B* can be seen most easily by substituting *A1+A2* for *A* in the above discussion, corresponding to subcomponents *A1* and *A2* in series, and calculating the *B* of each.

If performance changes in such a way that

$$A \rightarrow A + \Delta A,$$

then

$$CDF \rightarrow CDF + \Delta CDF = (A + \Delta A) * X + (A + \Delta A) * Y + Z$$
$$\Delta CDF = \Delta A * (X + Y) = \Delta A * B(A) = \Delta A * FV(A) * \frac{CDF}{A}.$$

The last line shows how to estimate changes in *CDF* without re-solving the entire model, once the "importance measures" *B* and *FV* have been obtained. This is basically how the MSPIs are evaluated. The above example can also be used to illustrate certain pitfalls. In practice, computations of importance measures are usually based on truncated cut set expressions. Suppose, for example, that *Y* is small, and that the contribution *A*Y* is accordingly "truncated" from the model in the process of obtaining the *CDF* expression. Then the contribution from *X* will be reflected in the *FV* and the *B*, but the contribution from *Y* will not be reflected in the *FV* and the *B*. If *Y<<X*, this is not an issue, but when discrepancies appear in different estimates of *FV* and *B*, truncation is one place to look. Sensitivity analyses have found that cut-set truncation limits in the PRA models should be set at about 1E-12/yr or lower, to ensure acceptable accuracy in the calculation of FV.

The change ΔA is quantified as the difference between the current estimate of *A* and a "baseline" estimate of *A*. This will be discussed in the following subsection.

The factor *B(A)*, the coefficient of ΔA in ΔCDF, is in essence the context of *A*. If it is overestimated, then the change in *CDF* will tend to be overestimated, and vice versa. The accuracy of current and baseline estimates of *A* are likewise important. For example, if the baseline is underestimated, then adverse changes in *CDF* will be overestimated, while if the baseline is overestimated, then adverse changes in *CDF* will be underestimated.

Caveats

The actual calculation of the Birnbaum for certain elements may be a bit more complicated than implied above, partly because some software packages work by post-processing a Boolean equation for *CDF* that is lacking certain information. Examples of post-processing include how recovery actions are treated and how certain configurational restrictions are addressed. For example, software packages that delete concurrent maintenance actions may behave unexpectedly if one unavailability is driven to unity (as suggested in the definition of Birnbaum offered above): it might force the other unavailability to zero. There could also be some question

of how common cause failure (CCF) ends up being treated, when the basic event probability that drives the CCF evaluation is taken to unity.

Some of these considerations can be addressed by adding a small ΔA to A, computing the resulting change in CDF, and developing $B(A)$ from that.

3.3 Estimation of ΔA

ΔA, corresponding either to a change in UA or a change in UR, is estimated as the difference between two quantities: the baseline estimate, and the estimate of current performance derived from recent performance data. Derivation of baseline values is the subject of much work discussed elsewhere. According to NEI guidance, the estimate of current performance should be done along lines discussed in NUREG-1753. For each UR parameter, current UR data are used to update the prior distribution for that parameter, and the estimate of current performance is taken to be the mean of the posterior distribution.

Per NEI guidance, the prior used for UR is the "constrained non-informative prior" (CNIP) (refer to NUREG-1753). This prior is "constrained" to have a mean value equal to the baseline, and a spread that is "non-informative" in a mathematical sense that is beyond the scope of this summary overview. In essence, the CNIP has the desirable property that relatively sparse data can quickly push the posterior distribution in the direction of the data, but the posterior mean is still less volatile than maximum-likelihood estimates (i.e., the posterior mean fluctuates less than estimates derived simply by dividing failures by demands).

The selection of this prior distribution was discussed at some length in NUREG-1753. Through examination of selected examples, this approach was shown to be preferable to the following alternative decision rules: (1) using the maximum likelihood estimate, and (2) using a prior that is centered at the industry mean and having a spread corresponding to uncertainty in the industry mean.

3.4 Definition of MSPI's

As currently formulated, the MSPI of a system is a simplified and linearized approximation to the change in CDF due to changes in reliability and availability of risk-significant elements of that system. The calculation focuses on key components, and quantifies the change in CDF using a simple formula based on importance measures.

The MSPI is formulated as a sum of changes related to UA and changes related to UR:

$$MSPI = UAI + URI$$

Unavailability-Related Contributions

UAI, the UA-related contribution, is a sum of contributions from different trains:

$$UAI = \sum_{j=1}^{n} UAI_{tj}$$

The summation runs over trains, and UAI_{tj} is the contribution of the jth train to the change in CDF due to changes in unavailability of this train.

If contributions to a given train's unavailability can be collected into a single PRA basic event having unavailability UA_t, then (see the above summary regarding importance measures) the change in CDF associated with a change in train UA can be written as (Ref. 2)

$$UAI_t = B(UA) * \Delta UA$$
$$UAI_t = B(UA) * (UA_t - UA_{BLt})$$
$$UAI_t = CDF_p \left[\frac{FV_{UAp}}{UA_p} \right] (UA_t - UA_{BLt}),$$

in which items carrying a "p" subscript are understood to be calculated using the "P"RA values, while items on the right-hand side *not* carrying a "p" subscript (carrying instead a "t" subscript) are derived either from current operating data or from baseline data. (In the NEI formulation, the "t" subscript just refers to "train.") This is an important and useful point: this formulation divorces the calculation of *B(UA)* from the calculation of ΔUA. Per the earlier discussion, *B(UA)* is independent of the value of *UA*, and basically furnishes context for *UA*. Given *B(UA)*, the terms whose difference yields ΔUA need only to be calculated on a mutually consistent basis – not necessarily consistently with the PRA - in order for the formula to yield a good estimate of the change in CDF. Of course, if *CDF* and *FV* are calculated and combined as above, then in order to yield *B(UA)* as desired, *CDF* and *FV* both need to be based on the same value of *UA* that appears in the denominator of the formula.

In practice, UA data are collected on a train basis. This avoids the potential for the overestimation of train unavailability that could result if individual components' unavailabilities were collected and summed as if they were independent.

The Unreliability-Related Contribution

The treatment of the UR-related contribution generally follows the above treatment of UAI. However, the elemental contributions to train unreliability need to be assessed separately, and partly as a result of this, there are additional considerations in URI.

One could imagine writing URI as a sum of contributions from each item in the system:

$$URI = \sum_{i=1}^{n} B(i) * \Delta UR_i ,$$

where i indexes components. For several reasons, this choice has not been made. Possible reasons include the following:

- Similar components within a given segment whose recovery and whose common cause are treated similarly ought to have the same B, but truncation effects may introduce spurious differences in the B values.

- Components having very low B values cannot contribute significantly to URI, but will clutter up the calculation, and would increase the burden of data reporting.

- Highly reliable components that normally experience no failures will clutter up the calculation, and may in fact introduce small negative contributions to URI whose meaning would need to be discussed if they were to be included.

Therefore, in an attempt to simplify and streamline the calculation, guidance has been formulated whose intent is to base the estimation of URI on components having significant B values and also enough failure potential to make them worth tracking. These conditions can be seen to be satisfied qualitatively by active components whose failure to actuate or otherwise change state can, *by itself*, cause failure (in the PRA sense) of an entire "train." Examples of elements *not* satisfying this condition are passive components, and components appearing in parallel with other components in the same train that are capable of fulfilling the mission (such as valves in parallel flowpaths associated with a single pump "train").

Defining "train" has been problematic. It is tempting to try to base the formulation of "train" on segments including pumps, but examples have been found in which leading system cut sets including active components are not captured by this definition. For example, if the discharge of several pumps is headered together, and there are fewer flowpaths downstream of the discharge header than there are pump "trains" upstream, then the flowpaths may yield the dominant cut sets. However, their elements would not be included in guidance that considered only the segments containing pumps.

There is no real technical issue associated with train definition. The generally-acknowledged purpose of the MSPI formulation is to focus effort on the real contributors, and the only question is how to articulate general guidance that will lead to the desired efficiencies without creating inconsistencies and confusion.

Currently, the guidance (Ref. 2) is to work with the following quantity:

$$URI = CDF_P \sum_{j=1}^{m} \left[\frac{FV_{URcj}}{UR_{pcj}} \right]_{\max} (UR_{Bcj} - UR_{BLcj}),$$

where

the summation is over those active components in the system that can by themselves fail a "train,"

CDF_p is the plant-specific internal events, at-power core damage frequency,

FV_{URc} is the component-specific Fussell-Vesely value for unreliability,

UR_{pc} is the plant-specific PRA value of component unreliability,

UR_{Bc} is the current estimate of ("Bayesian corrected") component unreliability for the previous 12 quarters,

UR_{BLc} is the historical baseline unreliability for the component.

Most of the structure of this formula is by now familiar: it is essentially a way of calculating and summing different components' $B(UR) * \Delta UR$. However, the "max" subscript on the brackets

requires comment. Recall that the Birnbaum of elements in series is, in principle, the same, apart from the caveats regarding how Birnbaum may actually be calculated in practice, and the fact that truncation may affect events unequally. The "max" means that in the *URI* calculation, the *B* should be calculated for each component in a group in a given segment, and then the maximum of those *B* values should be applied to all elements of that segment. Apart from the caveats in Section 3.2 and truncation issues associated with importance measures, events in series ought all to have the same Birnbaum anyhow; choosing the largest value is an apparent simplification. It may even be slightly conservative if, for some post-processing reason, the *B* values of different basic event types actually ought to be different.

In light of the above, it is not unreasonable to ask why not advocate the notationally simpler formulation in terms of *B*. Based on points made so far, this would simplify the discussion. However, some believe that available software is less uniform in its capability to address *B* than *FV*, and *FV* is certainly more widely used in risk-informed regulatory practice. Moreover, a later section will address treatment of common cause failures (CCF), and because of that issue, there appear to be technical reasons for keeping the formulation in terms of $CDF * \frac{FV}{UR}$. Specifically, there is a need to calculate a *group* importance measure in order to capture all relevant contributions including CCF, and for this purpose, it is more practical to work with *FV* than with *B*.

4. Benefits of MSPI's

1. The MSPI's treat UR along the lines of the treatment of UR in NUREG-1753. This treatment is based on failure and demand counts rather than fault exposure time, and is intended to resolve certain issues associated with the way in which existing PI's treat fault exposure time. (However, see "limitations" below.)

2. MSPI's are simple to calculate. An MSPI requires only baseline performance parameters that go into the priors, and a set of importance measures derived from a plant model. Once the importance measures are derived, manipulating the plant model is no longer necessary in order to quantify the MSPI. Given the above parameters, the MSPI can be quantified by hand calculation (although a spreadsheet will normally be preferred).

3. The MSPI rolls up most equipment performance data into a single performance-related figure of merit for each system. Therefore, although the MSPI addresses both reliability and availability, and spans non-diverse trains within a given system, the number of different performance indices that need to be addressed in an Action Matrix is kept to a minimum.

4. Despite being simple to calculate, the MSPI can be a very good approximation to the change in *CDF* due to current performance, provided that changes in performance are not extremely large, and provided that current performance can be estimated accurately. If the changes in performance are large, the MSPI's correspondence to change in *CDF* loses numerical accuracy (see discussion below), but the MSPI still points up the existence of a large change.

5. Limitations of MSPI's

NEI 99-02 implicitly notes limitations of the MSPI as follows:

> Due to the limitations of the index, the following conditions will rely upon the inspection process for evaluating performance issues:
>
> 1. Multiple concurrent failures of components
> 2. Common cause failures
> 3. Conditions not capable of being discovered during normal surveillance tests
> 4. Failures of non-active components

These and other limitations are discussed below.

5.1 Multiple-Failure Events

Implications of multiple-failure events are not fully addressed by the MSPI. The way in which a multiple-failure event enters the MSPI calculation is that the failures are scored as if they occurred in separate events. The risk calculation reflects common cause factors, so the potential for common cause failures is (or can be) reflected in the MSPI,[***] and common cause failure probability increases if the baseline unreliability increases. However, the common cause parameters themselves are not updated as a result of performance data. This treatment is arguably appropriate for events in which failures are known not to be correlated by a common cause or a cross-cutting performance issue. However, if observed failures are correlated causally, then the MSPI approximation may be non-conservative ("may be" because even at baseline performance levels, we expect common cause failures to occur occasionally). Risk-significant multiple failure events will continue to be treated separately from MSPI's.

5.2 Effect of Linearized Approximation

Because of the simplified calculational approach, the MSPI tends to understate the change in *CDF* when redundant elements have large increases in unreliability or unavailability. This point is illustrated in the following table, based on postulated changes in a hypothetical contribution to *CDF* from the product of two independent probabilities X and Y.

Table 1 Error in Estimated Changes in Contribution X*Y

% Change in X	% Change in Y	Actual % Change in X*Y	Linearized-Estimate % Change in X*Y	Error in Estimated Change
10	10	21	20	1/21
20	20	44	40	4/44
100	100	300	200	100/300

When the changes in X and Y are small, the error is insignificant.

[***] Some of the implications of CCF can be reflected in the MSPI by appropriately calculating the FV importance used in the MSPI formula (i.e., summing over CCF-related contributions).

5.3 Scope of Components Included in MSPI

For simplicity, only certain components are included in the MSPI calculation. Others are omitted for one or both of the following reasons.

- Their Birnbaum importances are low, so that only extraordinary performance changes would significantly affect the MSPI even if they were included.

- The number of failure events associated with them is relatively low, so that most of the time, including these components would either pointlessly clutter up the calculation or introduce a small non-conservative bias.

The latter comment warrants a brief explanation. According to the current formulation, for each included component type, current failure data need to be processed through the CNIP in order to obtain a current estimate of item performance. In every case, this introduces calculational steps. In the case of passive components, for which there would typically be no recent failures in the component pool for a given component type, inclusion of those items in this process would merely result in a very small negative contribution to the assessed change in *CDF*.

Passive components are a special case of components not included in the MSPI calculation. As noted in the NEI 99-02 quote above, activities other than MSPI will address these components.

5.4 Limitations of the Constrained Non-Informative Prior

Use of the CNIP has some limitations. The use of the CNIP is better than using no prior at all (e.g., taking a maximum-likelihood approach), but is, in principle, inferior to using an actual, current, state-of-knowledge prior regarding item performance. Using a state-of-knowledge prior would require a very significant analysis effort, so using the CNIP instead is a major saving in effort. However, for items having very high baseline reliability, the CNIP requires a large number of failures before its mean value shifts very much. This behavior arises because a very small unreliability implies a significant body of prior experience, which is clearly appropriate in a prior if we wish to estimate a parameter based on pooled data. However, in the present case, we are trying to estimate a *change*. The CNIP has been shown to behave reasonably well except when the prior mean is very small.

6. Treatment of Fault Exposure Time in SDP and MSPI

Recent comparisons between MSPI and SDP treatments of operational events have shown that in some cases, there are significant differences between the results of the two treatments. The purpose of this subsection is to illustrate some of the reasons why these differences occur, and suggest how to resolve them. This is related to the point noted in NEI 99-02 and quoted above, that conditions not discoverable in a surveillance test are not fully addressed by the MSPI.

The following table presents a simplified comparison between the SDP and MSPI treatments of current operating experience.

Suppose that during a 3-year observation period of the MSPI, an inspection finds an instance of deficient performance associated with equipment UR or UA. The resulting SDP performance band assignment can be compared with an MSPI performance band assignment covering the time period within which the inspection finding occurred.

By design, the MSPI averages data over an observation cycle, and thereby addresses longer-term performance issues. The SDP does no such averaging; it assesses the risk contribution from each instance of degraded performance. Because the MSPI averages data and the SDP does not, it is easier in some ways for a single failure event to lead to a non-green finding under the SDP than under the MSPI process. The interpretation is correspondingly different: the MSPI implies a change in CDF, while the SDP finding relates to an episodic risk contribution.

Given that a component is found to be in a failed state, the two approaches differ significantly in how they treat the time period during which a component was (in retrospect) in that failed state. The SDP assesses unavailability for this time period, and quantifies its risk significance using calculations somewhat resembling those of the Accident Sequence Precursor program. In contrast to this, the MSPI uses the fact of the failure event to requantify the "demand" failure probability, and assesses the resulting change in risk. The MSPI approach is intended to capture, in an average sense, the time during which the component is in a failed state; more demands during an observation period lead to a larger denominator and a lower resulting UR value. However, the degree of success in doing this depends on details of the formulation, as illustrated below.

Table 2 Comparison of SDP and MSPI Treatments of Operating Experience

	SDP	MSPI (as currently formulated)
Purpose	*Estimate* the increase in CDF due to conditions that contribute unintended risk increases caused by deficient licensee performance as *observed* in inspections	*Infer* deficient licensee performance by *detecting* increases in CDF as manifested by changes in model parameters
Input	Inspection Findings related to SSC Unavailability	Unreliability (UR) data, unavailability (UA) data, event frequency data
Role of exposure time	Exposure time (duration of the discrepant performance) is used in the quantification	None (as currently formulated)
Pooling of data	None. Uses information from event or condition that is subject of inspection.	Failure data for like components are pooled, so that they are all assigned the same UR
Statistical treatment	None	Bayesian reliability with confidence limits
Observation Period	One year. When a condition is discovered, the SDP determines the exposure time, and evaluates the risk contribution from this episode normalized to one year observation period.	Three year observation period is used to determine estimate of current performance; chosen duration balances the need for good statistics against need to detect performance changes within a reasonable time.
Output	Conditional core damage probability associated with unavailability based on exposure time => Performance Band Assignment	Change in CDF due to change in underlying UR and UA model parameters => Performance Band Assignment

For simplicity, consider the case of a pump having two dominant causes of failure, numbered 1 and 2, each with its own standby failure rate (λ_1, λ_2 respectively). For simplicity, focus on a one-

of-a-kind component, so that issues of data pooling do not arise. Suppose that there are two kinds of tests of this component. One kind of test is performed monthly, and can detect failures of type 1; the other, a more stringent and more inconvenient test, is performed every 18 months, and can detect failures of either Type 1 or Type 2. (When a Type 2 test occurs, it supersedes the Type 1 test, so that in a given month, there is a single test.) Suppose further that "real" demands occur very rarely, so that we can neglect the renewal effect of actual challenges.

Then, using the usual approximations (neglecting repair time, and so on), we can write the contribution of these standby failure modes to unreliability, averaged over time, as

$$UR = \tfrac{1}{2} \lambda_1 T_1 + \tfrac{1}{2} \lambda_2 T_2 + \ldots,$$

where T_1 and T_2 are the test intervals (in the present example, 1 month and 18 months, respectively).

Although the present MSPI treatment of UR distinguishes "failure to load and run" from "failure to run after operating for a while," the present MSPI treatment lumps all fail-to-start causes together into a single parameter that is determined by counts of failures and demands. Consider now the apparent risk significance of 1 failure event within a 36-month observation period. One way of treating these data (corresponding essentially to the current MSPI framework) is to update the UR prior with one failure and 36 demands (34 Type 1 monthly demands and 2 Type 2 demands at 18-month intervals). Depending on the component-specific prior, this would tend to give a delta-UR on the order of 2 or 3 percent. (The MLE would be 1/36, and these data will pull the posterior mean towards that value to a degree determined by the spread of the prior.) On the other hand, treating these data within the SDP framework, we would first establish whether the failure was Type 1 or Type 2, and assess the UR contribution in terms of the time period during which the component was apparently in a failed state. If it is determined to be a Type 1 failure, and we find that the component was unavailable for one month out of the 36 months in the observation period, we would get an answer not too different from that of the MSPI treatment (depending on the MSPI's prior). Because the MSPI uses a prior, and works with a 3-year observation period, the MSPI-determined change in UR resulting from a single failure event will tend to be less than the SDP-determined change in UR for the same event. For this Type 1 example, the discrepancy is not too large; but if the failure is a Type 2 failure, and is discovered at the end of an 18-month period, the SDP-determined contribution to UR can be more like 50%, rather than 2% or 3%.

The MSPI treatment changes very significantly if we distinguish the two failure modes, and give each its own prior. That is, we could establish different priors for failure modes 1 and 2, and update them separately. Instead of writing

$$UR = p,$$

and updating with all data to get p, we write

$$UR = p_1 + p_2,$$

and separately update p_1 and p_2. If the only failure in the observation period is Type 1, we update the Type 1 prior with one failure and 36 demands, and update the Type 2 prior with 0 failures and 2 demands. If the only failure is a Type 2 failure, we update the Type 1 prior with 0 failures and 36 demands, and update the Type 2 prior with 1 failure and 2 demands. Depending

on the spread in the Type 2 prior, this treatment might yield a performance band assignment closer to the SDP's.

Discussion

Recent history shows that current operational norms can allow failures to go undiscovered for some time, even if there is no apparent programmatic deficiency from a conventional point of view. For a high-B component, such events have high apparent risk significance. It can be asked whether current operational norms are adequate, or how the regulatory response should reflect this consideration. Those kinds of issues are beyond the scope of the present discussion.

In the above example, applying the MSPI philosophy in more detail tends to yield a result that resembles the SDP result. This suggests a way of dealing with those differences between MSPI and SDP that derive from modeling approximations. However, complete equivalence of the two approaches is not necessarily desirable. They have distinct purposes. The SDP is carried out *conditional on* deficient performance; the MSPI is another way of *looking for* deficient performance. Correspondingly, a key element of the MSPI approach, not present in the SDP, is the use of a prior distribution. One reason to use a prior is to reduce the incidence of unwarrantedly negative conclusions based on sparse performance data. More broadly, the role of the prior is to incorporate a broader knowledge base into the assessment. Some differences between the two quantification approaches can be expected as a result of the use of the prior in the MSPI, even if a similar use of fault exposure time is made.

In order to play its proper role, the prior needs to correspond appropriately to prior knowledge. The MSPI prior would ideally reflect prior knowledge about the probability of performance issues leading to changes in UR. One could imagine basing such a prior on past history, including recent inspection findings. Presently, for "demand" failures, the MSPI uses essentially a single-free-parameter CNIP; the "a" value is predetermined, while the "b" value is established by the prior mean (taken to be the "baseline" value). No explicit consideration is given to the prior likelihood of performance changes; rather, the properties of the CNIP have simply been found to be better than using MLE in some typical examples. However, as noted elsewhere, the CNIP displays significant inertia when the prior mean is small, and this would need to be considered carefully before the CNIP was adopted for application to the kind of low-probability, few-demand events typified by Type 2 in the above example.

Both the more-detailed MSPI and the SDP end up requiring equivalent information for deciding how to bin failures of a given component. In the SDP, one establishes how long the component has (probably) been unavailable. In the hypothetical more-detailed MSPI posited above, one accomplishes a similar result by establishing failure mode type and relating this to test interval. If the condition is identified in a Type 1 test, the failure can be scored as Type 1; if identified in a Type 2 test, more investigation is needed. These are two complementary ways of getting at the same underlying point.

It has long been recognized that it is inappropriate to equate all failures of a given component. For example, an issue formally similar to this arose in the RBPI development, in connection with long-term failure-to-run events. In that case, the issue was resolved by distinguishing between failures based on when they occur within a trial. The present discussion is, in essence, a generalization of those earlier discussions.

In summary, the example suggests that properly formulated, the MSPI can implicitly address fault exposure time issues. Currently, conditions "not capable of being discovered during normal

surveillance tests" (as stated in NEI 99-02), corresponding to Type 2 failures in the above discussion, are outside of the scope of the MSPI, and therefore need to be treated within the SDP.

7. Key Issues Affecting MSPI Development and Implementation

7.1 Relationship Between Significance Determination Process (SDP) Outputs and MSPIs

Some operational data are currently being analyzed under the SDP. Because the MSPIs do not currently address concurrent failures, the significance of such events is currently within the purview of the SDP.

As illustrated in the previous section, fault exposure time is treated differently in the two approaches. One objection to the original UA-based performance indicators was the way in which they captured failure information, namely, that when a failure was discovered, UA was estimated based on the time since the component was last known to be good. In some cases, this leads to a very large UA contribution, whose appropriateness is currently being debated. This problem was addressed in NUREG-1753 by treating UR and UA separately, and basing the UR treatment on failures and demands rather than on a time-dependent reliability model. However, that approach to estimating UR required certain assumptions regarding whether all "demands" are really equivalent.

7.2 False Assignment Probabilities

Work in the Phase 1 RBPI development showed that UR indicators in particular can have significant false indication probabilities. This is especially true for items having a high Birnbaum importance and for which few failures are expected within an observation period. Under these conditions, the relative scatter in observed number of failures ($\Delta N/N$, where N is the number of failures) is large. This means that under some conditions (as presented in NUREG-1753), there can be a substantial false indication probability, especially at the green / white interface, where $\langle N \rangle$ can be small.

At this point, the analog of NUREG-1753's investigation of false indication probabilities for MSPIs has not yet been conducted. It is possible that combining the elements, as done in the MSPI, changes the picture from what was seen in the RBPI work. Frequently, current estimates of element UR will fall below the baseline (which is, after all supposed to be the mean); this will introduce negative contributions to the MSPI that can balance out positive contributions. Since the MSPI sums over fluctuating quantities, the relative statistical error in the MSPI ought to be less than in any of the constituent contributions, but the actual numerics of this have not yet been explored systematically.

It is nevertheless clear that significant *statistical* false-positive indication probability exists in some of the MSPIs.

As long as the MSPIs make use of the statistics of infrequent events occurring within short time windows, substantial false indication probability will remain for high-*B* elements. However, it is possible to improve the performance-band assignment process by embedding the MSPI evaluation in a more considered decision-theoretic framework, as discussed in Ref. 3. In such a framework, the chance of a false indication is weighed against the *consequences* of a false

assignment before the performance band is assigned. One possible result of such an assessment is that it would take more failures to go from green to white than it would if the white assignment were to be made based solely on the mean of the posterior distribution, without regard to the consequences of a false positive.

7.3 "Invalid Indicators"

"Invalid indicators" are MSPIs having the property that just one failure above baseline during the observation period can cause the index to go "white." According to Appendix F (Ref. 2),

> If, for any failure mode for any component in a system, the risk increase (ΔCDF) associated with the change in unreliability resulting from single failure is larger than 1.0×10^{-6}, then the performance index will be considered invalid for that system.

This is a special case of the false-indication issue discussed in the preceding subsection. The determination of "invalidity" is based on both the unreliability contribution to the MSPI (propagating the postulated one failure through the prior for *UR*, etc.) and the unavailability contribution (recognizing that there will be some unplanned unavailability associated with the failure, and per NEI draft guidance, quantifying it based on the assumption that the downtime is half of the tech spec AOT for that train).

Based on current draft NEI guidance, "invalid indicators" are not to be used. The potential downside of using them is easy to understand: a false white might occur someday if such an indicator is used. On the other hand, blanket exclusion of such indicators may not be the best choice.

The following possibilities for partially addressing this issue were discussed at a recent workshop.

- For URI, increase the data collection period from 12 to 20 quarters

  This would make the indicator somewhat less noisy, but also less responsive. Work would be needed to establish a significant benefit in reduction of false-positive probability, and to show that the increase in false-negative probability would be justified.

- Expand the component population of like types (across systems and/or across units) to improve the statistics of small numbers

  This would also make the indicator less noisy, at the cost of more data collection. It would also need to be clarified that the risk significance of the newly-added components would not contribute to the URI index. If the component population is in fact homogeneous – if all components in the population are subject to the same performance influences – then increasing the population of scored components increases the accuracy of the indicator. If the risk-significant components are affected differently from the population as a whole, then the URI is correspondingly biased.
- Change the *Prior Distribution*

  The prior can be changed to make a white less likely. The prior currently used, the CNIP, simply damps out some of the statistical noise. It could be changed to damp out more of

the noise. Currently, the decision rule goes white when the mean of the posterior goes white, even if there remains a substantial probability that performance is still green. Therefore, solving the "invalid indicator" problem with the prior alone (leaving the rest of the decision rule as-is) would mean that the mean of the posterior would have to be biased more towards green.

An alternative is to pay more attention to the residual "green" probability, and adjust the thresholds as described below.

Improving the prior (not just biasing it to green) would be one element of a more decision-theoretic approach. For this application, a prior should ideally reflect a state of knowledge distribution regarding the potential for performance issues to degrade UR.

- Adjust the thresholds to ensure low probability of false positives and false negatives

  "Adjusting the thresholds" to optimize false-diagnosis probabilities is a traditional application of statistical decision rules. Here, "adjusting the thresholds" does not refer to changing the 1E-6 / 1E-5 / 1E-4 approach to defining performance bands. Rather, it refers to *how many failures would be needed* in a particular index before the probability of a true "white" (i.e., DCDF actually > 1E-6) would be deemed sufficiently high to warrant that performance band assignment. A statistical decision rule can optimize this number-of-failures threshold as a function of the prior probability of a performance excursion, and the rewards and penalties for true and false diagnoses (Ref. 3).

- Identify the components with "Invalid Indicators" and use statistical tests of adverse trending rather than the URI and UAI measures

  One way of stating the problem with invalid indicators is to point out that one failure is not a "trend." Accordingly, one way of dealing with performance areas whose MSPI's are considered "invalid" is to frame the question as one of whether recent failures (presumably more than 1) appear to correspond to a trend. To the extent that calendar time influences this evaluation, it has the potential not only to address "invalid indicators," but also to address the situation in which a seemingly significant combination of failures occurs within a short time. In this latter hypothetical situation, it may be the case that significant statistical evidence of an adverse trend is present, even though the number of failures observed is not sufficient to trip the MSPI. Rather than waiting out the 12 quarters, it would be desirable to react sooner. Applying statistical trending methodology to both of these issues has benefits, and will be explored.

Also, the use of more up-to-date baseline UR values results in fewer invalid indicators. This phenomenon needs to be investigated further.

7.4 Effects of Using the Constrained Non-Informative Prior (CNIP)

The CNIP was introduced in the Phase-1 RBPI Development because its use in performance-band assignment improved the false diagnosis probabilities, relative to other candidate priors evaluated at that time. Although not based on an explicit assessment of our state of knowledge regarding the variability in licensee performance, the CNIP at least has the desired effect of smoothing out statistical fluctuations, up to a point. For selected examples, the CNIP's performance was compared in NUREG-1753 to (a) using a maximum likelihood estimate (dividing current failures by current demands) and (b) using a prior whose spread was

representative of uncertainty in the industry mean, as opposed to variability in a specific plant's performance. The CNIP out-performed both of these alternatives for the examples analyzed. (The MLE based on recent data is too noisy, while deriving an estimate from a prior based on uncertainty in *industry* performance is not sensitive enough to current plant-specific performance.)

However, the CNIP has one feature that is a drawback in some cases: for highly reliable items, the CNIP has very substantial inertia. The mean of the prior is fixed at

$$\frac{a}{a+b},$$

with $a \approx \frac{1}{2}$, so that when unreliability is very small, b is very large. This means that when the prior is updated (by adding failures to the numerator and demands to the denominator), many failures are needed in order to move the mean of the posterior significantly away from the mean of the prior. It is not clear that this property is desirable in a methodology aimed at responding to performance trends. This warrants further exploration.

7.5 Treatment of Common Cause Failure (CCF)

It appears that both indicators and inspections were originally intended to address CCF (SECY 99-007, Ref. 4). However, the current draft industry guidance for MSPIs (Ref. 2) states that

> Some aspects of mitigating system performance cannot be adequately reflected or are specifically excluded from the performance indicators in this cornerstone. These aspects include ... the effect of common cause failure, ...

The industry approach appears to relegate regulatory oversight of common cause potential entirely to inspection processes. Given a CCF-induced multiple failure, this will be analyzed under the SDP, but it is not clear that the program is intended to address, before the fact, the existence of conditions that promote CCF. In any case, it is arguably desirable to reflect the CDF significance of all performance changes that can validly be reflected in the MSPI, given the purpose of the MSPI and the character of the performance data and the available models. The present subsection discusses this issue.

Some CCF models represent the CCF contribution to risk as being essentially proportional to overall failure probability. In such models, if the measured UR increases and the proportionality constants are left alone, the assessed CCF contribution increases along with the independent failure contribution. This is how the NRC staff effort has approached MSPI quantification: a change in UR increases CDF both through the independent failure contribution and through a CCF contribution. The industry approach does not add the CCF contribution. For a given data set and a given model, the current staff approach therefore estimates a larger CDF change than does the industry approach. In some cases, this leads to lower number-of-failures thresholds.

It can be debated whether performance changes that increase UR typically, or ever, also increase the CCF contribution. To fulfill the intent of the MSPI, the industry approach is appropriate if the answer to this is "no." The current staff position is arguably appropriate if the answer is "perhaps."

One could consider addressing CCF within the MSPI by replacing the parametric models with a standalone model for the CCF contribution. Failure events could be analyzed to determine whether they manifested underlying common cause mechanisms, and corresponding CCF contributions could be assessed individually. This sort of event review would make the exercise of MSPI quantification significantly more arduous. It also implicitly changes the purpose of the MSPI, which is to flag the existence of potential problems, not to diagnose them.

Because the purpose of the MSPI is to flag *potential* performance problems based on operating experience, it seems most reasonable to propagate changes in observed UR through the parametric CCF model, and include the change in CCF contribution in the assessed change in CDF. If there is an underlying performance issue causing a real increase in UR, it may well relate to CCF anyhow.

Some of the time, the underlying performance issue will not relate to CCF potential, and in that case the present recommendation would appear to be conservative. However, current estimates of the common cause parameters are small numbers, and recent sensitivity studies have shown that while including CCF does change the number-of-failures thresholds somewhat, the effect is not major (assuming that the common cause model parameters are set at current levels). Debating the accuracy of the common cause parameters is beyond the scope of the present report.

7.6 Validation

As was the case with RBPIs, there is a fundamental problem with "validating" MSPIs, if "validation" is construed as requiring that MSPIs be calculated for plants that are "known" to be in certain performance bands in certain areas. However, elements of the process can be checked for apparent reasonableness. In order for the MSPIs to behave correctly, the following need to be validated individually.

Baseline Performance
The baseline needs to be reasonable. Choosing baselines is a continuing topic of discussion. Note that choosing the baseline for UR elements not only fixes the negative term in the delta, it also influences the behavior of the CNIP. This introduces subtleties that are beyond the scope of this paper.

Estimate of Current Performance
For UR, the estimate of current performance is derived from the CNIP and current operating data. The basis for this still needs discussion. The fact that many of these estimates lie below baseline introduces negative contributions to the MSPI, a fact that calls for some reflection.

Estimate of Birnbaum (or CDF*FV/UR or CDF*FV/UA)
The Birnbaum must also be correct. As the table top exercises proceed, and SPAR Birnbaums are compared with industry Birnbaums, a certain amount of model reconciliation may need to occur.

The program is conducting "table top" exercises with industry, in which recent performance data are being processed with current models to derive MSPI values, which can then be judged to see whether the process appears to be behaving properly. The MSPI verification effort includes several tasks: evaluation of FV/UR and FV/UA values from SPAR models and comparison with corresponding plant PRA values, identification and resolution of differences, identification of

invalid and insensitive indicators, and evaluation of the sensitivity of results to assumptions concerning baseline values. These tasks are being performed for each of the 20 plants within the pilot program. Preliminary results indicate that SPAR model FV/UR and FV/UA results can differ significantly from plant PRA results. Therefore, an important part of the verification effort involves resolving these differences.

8. Summary

The MSPIs provide reasonable estimates of changes in CDF that are caused by performance changes that affect the UR and UA of risk-significant equipment within the scope of the calculation. Certain items are not amenable to being tracked with the MSPI, and are addressed separately.

The MSPIs offer the advantages over ROP performance indicators that were identified in previous work on candidate RBPIs, while avoiding certain disadvantages of the RBPIs. The advantages include separate treatment of UR and UA, allowing for more accurate modeling of contributions to UR (including the potential for improved treatment of fault exposure time), and plant-specific thresholds. One disadvantage of the candidate RBPIs was the proliferation of indicators, which resulted from the separate treatment of UR and UA, and the need to distinguish trains of different types. The proliferation of indicators creates issues of interpretation in implementation: it complicates the application of the Action Matrix. This proliferation issue is addressed in the MSPIs by rolling up all contributions for a given system into a single index for that system, addressing both UR and UA for all trains. In addition, the simplified (importance-measure-based) approach to calculating the MSPIs provides a good estimate of CDF changes due to the components covered, unless very large changes in UR and UA occur.

Certain issues need to be addressed before the MSPIs are finalized. The path to resolution of these issues is straightforward. Issues relating to the formulation of the MSPIs include the following:

- the treatment of CCF in the MSPIs,
- the adequacy of the demand-based UR model in dealing with fault exposure time, and
- the need for an improved way to deal with the false-positive issue.

Suggestions for resolving these issues were provided above, and are currently being explored.

In addition to the above issues, it is necessary to be sure that the Birnbaum values used to quantify the MSPIs are sufficiently accurate. (Per the earlier discussion, the Birnbaum of an item can be related to the FV, the CDF, and the item UR or UA as appropriate, but the discussion is most accurately carried out in terms of the Birnbaum.) The Birnbaum of a given item depends on what functionality is credited as being redundant to that item, how that functionality is modeled, what data are used to quantify the model, and (to some extent) the truncation value used. Work is underway to assess the Birnbaums by comparing industry models with SPAR models, and understanding and reconciling the differences.

9. References

1. Nuclear Regulatory Commission (U.S.) (NRC). NUREG-1753, "Risk-Based Performance Indicators: Results of Phase 1 Development." NRC: Washington, D.C. April 2002.

2. Nuclear Energy Institute (NEI). NEI 99-02 (Draft Report), "Regulatory Assessment Performance Indicator Guideline," Section 2.2 ("Mitigating System Performance Index") and Appendix F ("Methodologies for Computing the Unavailability Index, the Unreliability Index, and Determining Performance Index Validity"). NEI: Washington, D.C. 2002.

3. Youngblood, R.W. and C.L. Atwood. "Alternative Decision Rules for Assessing Performance," Proceedings of the International Topical Meeting on Probabilistic Safety Assessment (PSA '02). American Nuclear Society (ANS), Detroit, Michigan 6-9 October 2002. ANS: LaGrange Park, Illinois. 2002.

4. Nuclear Regulatory Commission (U.S.) (NRC). SECY 99-007, "Recommendations for Reactor Oversight Process Improvements." NRC: Washington, D.C. 1999.