On defining a 'hurt' metric to measure dissatisfaction with products

Divya Prasad, Martin Atkins, John McDermid & Ian Wand

Abstract

In industrial practice, complex, multiple criteria decisions are often based on the aggregation of information into a single number, without checking its theoretical validity. This paper considers four factors that contribute to the degree of dissatisfaction, or hurt, from the maintenance of products in service. The goal is to synthesise them into a single metric that can be fed back into the development and maintenance lifecycle. We outline a step by step procedure for constructing such a metric, fulfilling necessary theoretical obligations. We show that overlooking any step in the procedure can lead to flawed metrics. The lessons learned are applicable to any single metric intended for making multiple criteria decisions.

1. Introduction

A major aerospace company was interested in devising a 'hurt' metric, which they described as:

> *… A single metric that could convey the amount of dissatisfaction or hurt, that a product is giving an operator due to its unreliability, difficulty in testing and maintaining, cost of repair etc. …*

The problem owners proposed four metric drivers (contributing factors or criteria), that, in their experience, most affected the level of hurt. The purpose of synthesising these factors into a single number representing an overall score was to be able to compare different products (or components) functioning on the same aircraft, or similar products functioning on different aircraft. The components could be of varying criticalities and specifications. The metric could then be used to guide the collection of maintenance related data and to direct design and maintenance attention to where it is most needed.

Products (systems, subsystems or components) can be compared with each other by considering all the criteria one by one. In many cases, this is not enough to decide which product is the 'best' or 'worst' overall, or even to order or rank any two products. The reason is that no product may be better or worse than another in terms of all the criteria. The only way to resolve these conflicts and to arrive at a decision is to make trade-offs between the levels of achievement of different criteria. Multiple criteria decision making [1,2], deals with the articulation of such trade-offs to arrive at justifiable and rational decisions.

The method proposed for combining the four hurt metric drivers was a multiple criteria decision making method known as the Analytic Hierarchy Process (AHP) [3], which uses a linear combination to generate a single number. We were asked to assist in constructing the metric and this paper presents the results of our investigations.

We first review the proposed drivers (hereafter synonymous with criteria) and then outline a procedure that must be followed when constructing any such metric. In applying this procedure to the hurt metric we discover that some theoretical obligations are violated by the proposed set of metric drivers. The procedure also helps us overcome the violations and construct a simple and well-founded metric. The paper concludes with comments on how to apply this approach to the construction of other similar metrics.

2. The Hurt Metric Problem

The metric was initially motivated by the need to compare two different controllers for the same piece of avionic equipment. The following criteria were chosen by the company as the most important drivers for the hurt metric (with the intention that others would be added later):

SC: **S**afety **C**riticality (or Operational Effect)

MTBUR: **M**ean **T**ime **B**etween **U**nscheduled **R**emovals (or Maintenance Period)

NFF: Number of arisings resulting in **N**o **F**ault **F**ound

TRT: **T**otal **R**epair **T**ime

The term ‘arisings’ above refers to the number of times that a component was identified as the source of the current problem at ‘1st line’ (in-field). Once this happens the component must be removed (and replaced by a spare if necessary to continue operation), and sent on to ‘2nd line’ (a testing centre) for a confidence check before it can be refitted to the aircraft. If the 2nd line tests are not able to find any fault, the item’s NFF count is increased by one.

The following form of operational data and product design information was available for the two controllers initially under consideration (referred to as Item1 and Item2 below):

> *Over a period of* **h** *flight hours of Aircraft-Type* **t**, *there were* **a** *arisings, of which* **n** *resulted in no fault found. Total number of maintenance man-hours at 1st line was* **m**. Item1 *and* Item2 *are both safety-critical. The flight hours for* Item 2 *cover the whole fleet of aircraft, but not all these aircraft were fitted with the* Item2 *(so the data may be optimistic).*

The rest of the paper addresses the problem of generating a decision making model (a function) that combines the data for the drivers (the independent variables) into an overall value for the hurt (the dependent variable) caused by a component. The numbers returned by this function constitute the hurt metric.

3. Five-Step Procedure for Generation of Metric

The linear combination method used by AHP is only a particular instance of a more general method for combining multiple criteria ‘additively’ (where the contributions to the overall value are added in some way). There is a large body of work on the scientific principles of measurement [4,5], and multiple criteria decision making [1,2], which contains, among other things, the theoretical obligations that are imposed by the additive method. Based on a detailed study of this literature, we have developed [6] a step-by-step procedure that must be followed in the construction of any multiple criteria metric. Typical questions that must be asked and answered at each step and the results required are outlined below:

Step 1. Identification of Decision Alternatives for Comparison and Evaluation

Questions: What are the alternatives (entities) being compared? Are these already known or do we have to create them? Is the set of alternatives fixed or does it vary (with the criteria)? In other words, what is the domain of discourse?

Result: The set of decision alternatives being compared and evaluated.

Step 2. Definition of Evaluation Criteria and their Scales of Measurement

Questions: What are the criteria that will be used to choose between the alternatives identified in Step 1? For each criterion, is there a scale of measurement (a homomorphic mapping) [5] that can represent our empirical and intuitive understanding of the criterion?

Result: The set of criteria, with appropriate scales of measurement defined on them.

Step 3. Modelling of Preferences between Alternatives for each Criterion

Questions: What are the decision-maker's preferences between the alternatives based on each criterion (considered in isolation)? For each criterion identified in Step 2, is there enough information to say clearly that one alternative is '*preferred*', or '*equal*' to another, or are they '*incomparable*'? Do the preferences increase or decrease monotonically with increasing values of the criteria?

Result: A preference structure (set of binary relations over the set of alternatives) for each criterion.

Step 4. Application of Representation Theorem for Proposed Value Function

Questions: Does there exist a utility or value function [1,4,5] (e.g. of an additive or polynomial form) that can combine the criteria into an overall value? Do the preferences from Step 3 and the application specific trade-offs between the criteria fulfill the necessary and sufficient conditions of the representation theorem for the chosen value function (e.g., three or more criteria can be combined additively if, and only if, a condition known as 'mutual preferential independence' holds)?

Result: Chosen value function along with validation of its representation theorem.

Step 5. Construction of Multiple-Criteria Decision Model for Metric

Questions: Given the form of the value function chosen in the Step 4, what are the parameters needed to completely determine the value function? For example, in using the additive value function we must determine each criterion's contribution to the overall value. Various practical techniques (for example, AHP) are available to help with the construction of the final metric and to perform sensitivity analyses.

Result: Parameters to complete the model and hence the metric definition.

This five-step procedure must be iterated (refined locally or revisited in the light of insights from later steps) until a satisfactory metric can be derived. In some cases we may only be able to conclude that a proposed metric is infeasible. The specific characteristics of the application problem determine which of these steps are trivial and which are hard.

4. Identification of Decision Alternatives (Step 1)

Initially, Item1 and Item2 are the decision alternatives. This defines the domain of discourse of the problem – the set of possible alternatives are all aircraft components of this type. However, the original intention was to extend the set of alternatives to cover many more products on an aircraft. Although the alternatives would all be avionics equipment, the manner in which different components cause dissatisfaction differs (in terms of the criteria, their relative importance and their contributions to overall hurt). Thus, a broad range of alternatives may be too unwieldy to evaluate with the same set of four criteria. Here, we proceed with {Item1, Item2} as the set of alternatives.

5. Definition of Evaluation Criteria (Step 2)

Although four distinct criteria (SC, MTBUR, NFF and TRT) were proposed, no specific definitions were provided. This forced us to infer the exact meaning of the criteria from three sources: (1) the "usually accepted" definitions of the terms (2) the available data and (3) the issues and concerns raised by the owners of the problem. This situation indicates a poor application of Step 2 of the decision-making procedure, which will in turn make preference modelling (Step 3) and representation (Step 4) more difficult. The following sections analyse each of the drivers in detail.

5.1. Safety Criticality or Operational Effect

A single driver SC was intended to represent the criticality of a component to system safety or operation. The motivation for this is to be able to compare safety-critical, (other) critical and non-critical components with each other. The proposal was to use the 00-56 [7] classification of Safety Integrity Levels S1-S4 as a measure of Safety Criticality.

5.2. Mean Time Between Unscheduled Removals or Maintenance Period

This criterion also puts together two separate (although related) concepts. A distinction between unscheduled removals and scheduled maintenance-related removals is needed because the former gives more hurt than the latter. The current data also does not distinguish between 'major' and 'minor' removals (repairs) – this sort of classification may be needed to get an accurate measure of hurt due to frequent removals. The data provided on Items 1 and 2 did not make this distinction, so we elide it for now and assume that both types of removals contribute equally to hurt.

5.3. No Fault Found Rate

This is the number of times that no fault could be found at 2nd line for an item that was designated faulty at 1st line. A high NFF rate could indicate that 2nd line testing procedures are inadequate, or that better diagnostics are needed at 1st line. Thus, although the definition of NFF is clear and it has a simple measure (in terms of a count or a fraction), we will see in Step 3 that using NFF as a criterion is problematic.

5.4. Total Repair Time

Since no separate parameter is available to distinguish between the time spent dealing with an NFF occurrence and an actual repair, we assume that this driver represents the total time spent in testing or repairing the component. It is important (for fairness of comparison) to ensure that the data is obtained under identical conditions (using the same benchmarks), and is normalised to remove any bias from differences in sample size. For example, TRT should be taken only for those aircraft that are fitted with the item under consideration, not for the whole fleet.

6. Modelling of Preferences (Step 3)

In this step, the metric designers must make explicit what their preferences are between different alternatives that achieve different levels on each criterion. This could be done for each criterion in isolation by considering every pair of alternatives, and expressing a binary preference relation. This elaborate process can in many cases be shortened to a single statement, which says, for example, that the preferences increase monotonically with increasing values of a criterion. We model the preferences along the four hurt metric criteria below.

6.1. SC

The safety criticality of a component does not cause any hurt until the component fails: there is no inherent hurt in being safety-critical. This is very different from the other criteria that contribute directly to hurt. In fact the value of SC only changes the extent of hurt from all the other criteria. Thus, although the hurt increases with increasing values of SC, it seems unfair to penalise a high criticality component by simply ascribing a higher hurt metric to it than to a component of lower criticality, since, it not the component's fault that it is critical! We will see in Step 4 that the interaction between SC and the other criteria becomes significant in determining the validity of an additive combination.

6.2. MTBUR

The higher the value of MTBUR, clearly, the less the hurt. Thus preferences decrease monotonically with increasing values of MTBUR. In order to make the Step 4 easier, it is advisable to change MTBUR to a positively oriented criterion. We could take the reciprocal of the actual MTBUR, or the difference from some maximum possible value of MTBUR (say for a hypothetical perfect product). For the rest of this paper we assume that the label MTBUR stands for a suitably defined positive criterion.

6.3. NFF

The fact that no fault was found ought to represent good news, yet, ironically, there is a large adverse effect arising out of this apparent benefit, because the system has been temporarily unavailable, and there is uncertainty about whether or not the system continues to be faulty.

There are several ways in which NFF hurts: it represents a waste of time for both 1st and 2nd line staff. We assume that this time aspect would be folded into the driver TRT. However, an NFF occurrence means that the original problem has not yet been fixed. If the problem is intermittent then this could be potentially even more harmful than if it is persistent. If the component in question is critical, the hurt caused from a high NFF rate is much higher than if it is non-critical. A non decreasing NFF rate (per hour of operation) could also mean that 1st line staff have a poor understanding of the system's behaviour, which could be due to the system's complexity or simply inadequate staff expertise, but we should question whether the component should be penalised for this.

In conclusion, it is not clear whether we always prefer a lower NFF value, to a higher one, and this means that we do not understand how to model preferences for this criterion. In practice, we would have to return to Step 2 and redefine NFF or replace it with one or more other criteria that capture its intended meaning. For the rest of this paper, in order to proceed with demonstrating the remaining steps of the decision making procedure, we assume a simple preference model: that hurt increases monotonically with increasing values of NFF, which is the normal industrial/commercial interpretation.

6.4. TRT

It is clear that greater TRT values are more hurtful, however the total time spent in repair cannot be used directly, since the number of operational hours of the systems must be taken into account. The original proposal was to use total repair time per arising, but we replace it with total repair time *per operational hour* in order to compare aircraft that have different usage statistics.

We now proceed to Step 4, assuming that the preference models for all four drivers (SC, MTBUR, NFF and TRT) are of the kind "the higher the value of the criterion, the more its contribution to hurt".

7. Application of Representation Theorem (Step 4)

The representation theorem for using the additive value function to combine three or more criteria essentially states that:

*An additive value function exists **iff** the criteria are mutually preferentially independent*

The intuitive meaning of mutual preferential independence between criteria is that the preferred combination of any subset of criteria values (due to the relative importances assigned to those criteria) should be independent of the values of the remaining criteria. In other words, the contribution of any subset of (one or more) criteria to the overall value (of hurt) should not depend on the current contributions made by the other criteria. This is not true in the case of the proposed hurt metric, because as we have seen in Section 6, the relative importances of the three criteria MTBUR, NFF and TRT clearly depend on the value of the remaining criterion SC. This type of dependence between SC and the other factors, can be represented by the formula:

$$(SC_{low}, MTBUR_1, NFF_1, TRT_1) \geq (SC_{low}, MTBUR_2, NFF_2, TRT_2)$$
$$\not\Rightarrow$$
$$(SC_{high}, MTBUR_1, NFF_1, TRT_1) \geq (SC_{high}, MTBUR_2, NFF_2, TRT_2)$$

Here, '$\geq$' means "*preferred or equal to*", and the formula asserts that if, at a fixed level of criticality (SC_{low}), we prefer one combination of {MTBUR, NFF and TRT} above another, then, for a different fixed level of criticality (SC_{high}), it is not guaranteed that we still prefer the same combination. This implies that the subset {MTBUR, NFF, TRT} is not preferentially independent of its complement {SC}, and hence, no additive value function exists for combining these criteria. The preference relations in the other subsets do not have to be investigated further because a single preferential dependency is adequate to rule out an additive combination.

This analysis suggests that the contribution to hurt from the criticality of the component should be in the form of an external multiplier, rather than as a criterion.

8. Construction of Model and Metric (Step 5)

It is easy to proceed to Step 5, say from Step 2, and build an unsound decision model using the given drivers without checking the theoretical obligations of the intermediate steps. A naive linear combination of the four criteria could be used:

$$\mathtt{Hurt} = w_1.\mathtt{SC} + w_2.\mathtt{NFF} + w_3.\mathtt{MTBUR} + w_4.\mathtt{TRT}$$

where the w_i are constant weights as derived by AHP from judgements expressed by the decision-maker about the relative importance of the criteria. This shows how easy it is to create a metric that appears to be useful, but will provide misleading results because it does not correctly represent the preferences of the decision-maker. Hence incorrect decisions may result from skipping over some of the earlier steps and going directly to Step 5 of the procedure.

8.1. Decision Model without Criterion SC

As discussed in Step 4, the preferential dependence of three of the criteria on SC disallows a linear combination of the initially proposed set of criteria, and a simple solution would be to remove SC from the metric. This leaves a model that can only be used to compare items of similar criticality. Thus, the model has the form:

$$\mathtt{Hurt} = w_1.\mathtt{NFF} + w_2.\mathtt{MTBUR} + w_3.\mathtt{TRT}$$

For the chosen level of criticality we must derive the weights of the factors MTBUR, NFF and TRT using expert judgement. In order to compare items of dissimilar criticalities we could re-introduce the effect of SC, either as a multiplicative constant, or by having a model with a different set of weights for each level of criticality.

The theoretical validity of this smaller model (with only three criteria) can be verified by checking preferential independence between all the subsets of criteria and their respective complements. We do not pursue this further since the principles of the required analysis are the same as before and it is reasonable to assume that the remaining three drivers are preferentially independent for the purposes of this illustration. We can now proceed to feed operational data into the model and generate the values of the hurt metric for each alternative.

8.2. Generation of Hurt Metric for Item1 and Item2

The AHP technique is supported by a tool called Expert Choice [8] that helps to capture the decision maker's preferences between alternatives, and the relative importances of the criteria. The weights used in the linear combination can either be entered directly or are derived by AHP from relative importance judgements entered by a domain expert. Figure 1 shows the output produced by the 'performance sensitivity analysis' tool within Expert Choice. It shows the weight of each criterion (the vertical bars) and the contribution to hurt (the lines spanning the figure) calculated from representative operational data. The data itself is not shown here as it is entered elsewhere in the tool.

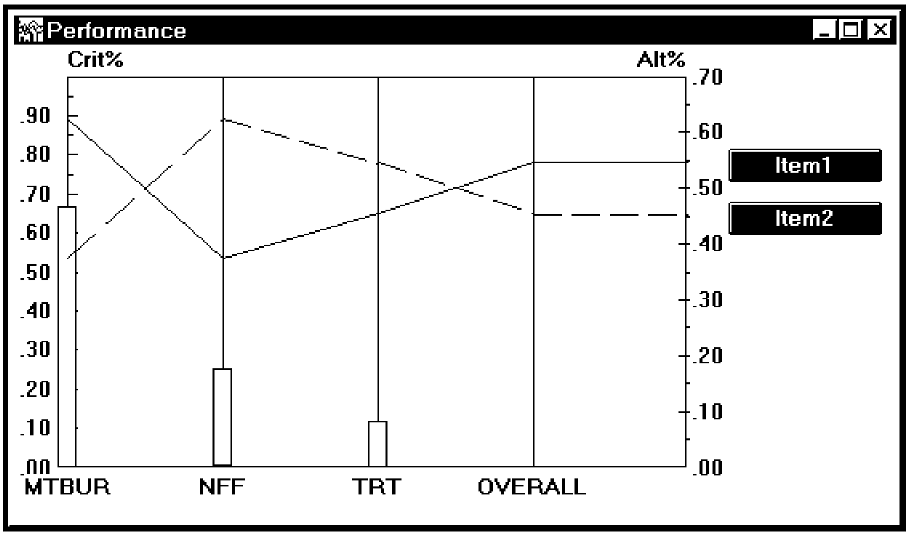

Figure 1: Example Weights and Results of Comparing Item1 and Item2

The weights can be adjusted by a domain expert depending upon the current application and the criticalities of the components under consideration. With the current set of weights, Item1 is more hurtful than Item2 in terms of MTBUR but less hurtful in terms of NFF and TRT. The overall hurt metric for Item1 (0.55) is higher than for Item2 (0.45), indicating that Item1 is more hurtful. For the same operational data, but a different set of weights, this result may change. The effect of component criticality can be incorporated into the metric by multiplication with numbers representing the impact of different levels of criticality. The models and metrics can be revised in the light of operational experience and improved understanding.

9. Discussion

It is not always feasible to produce a single number to measure multiple-criteria notions, yet we find that there is pressure from engineering industries to derive single numbers such as the hurt metric. If the derivation of such numbers is not well founded then any conclusions drawn from them may be invalid. The following pitfalls were discovered during the development of the hurt metric:

- **Ambiguous objectives**: It was not clear if the goal of the hurt metric was to interpret the available data, or to decide what kind of data should be collected. Furthermore, the definition of 'hurt' was itself unclear.
- **Ambiguous criteria**: Precise definitions for the drivers are an essential prerequisite for meaningful metrics to be generated. This is not surprising, as it only re-iterates the basic principles of measurement, however it is extremely important.

- **Preferential dependence**: An additive value function (e.g., linear combination such as that underlying AHP), should not be used when there is ‘preferential dependence’ among the drivers. The subset {TRT, NFF, MTBUR} is not preferentially independent of the subset {SC} because the relative (and absolute) importance of these drivers depends on the value of the driver SC.
- **No automatic checks**: Despite the theoretical invalidity of using the additive value function with the given criteria, there is nothing to prevent the practical construction of an additive ‘metric’. Indeed, construction methods (e.g. AHP) do not have formal checks built-in for verifying preferential independence. Once constructed, it is not obvious from an AHP model that the theoretical conditions might be violated.

Nevertheless, we were able to construct a simple and theoretically sound metric by eliminating the dependencies between the criteria. The resulting model allows items of similar criticality to be compared directly. It can be extended to deal with comparisons between items and their specifications, which would allow the notion of hurt to be refined into notions of ‘disappointment’ and ‘improvability’. This can be achieved by simply adding specifications and hypothetical ideals as elements to the set of alternatives in Step 1.

In some cases it may not be easy to verify the validity or otherwise of the theoretical obligations, and we might have to construct a model without being certain about meeting them. If such a model does not behave consistently with our empirical observations or expectations, it could be either because the model is theoretically unsound, or because it predicts results beyond our experience.

The validation of a model is tricky: on the one hand, the fact that the model can make predictions outside the experience of the decision-maker is a motivation for constructing the model; on the other hand, predictions that drastically disagree with the decision-maker’s expectations must cast suspicion on the model. Nevertheless, the use of a rigorous decision analysis procedure forces a discipline on the decision-making process for critical and complex decisions, which should lead to more justifiable, and better documented decisions.

10. Conclusion and Further Work

Industry often needs single numbers, such as the hurt metric, to help make decisions or to evaluate products, but does not have adequate procedures or tools to generate them in a well founded framework. We have seen that it is important to verify the theoretical obligations underlying any proposed multiple criteria metric. If these obligations are violated, we know from the theory, that the resulting decision models and metrics will be unsound, and may lead to decisions that contradict our understanding of the criteria.

The work described in this paper can be extended in the following ways:

- **Hurt Metric**: The need for synthesising a single number from the initial four drivers should be re-examined. The goals of the metric must be clearly stated and the criteria must be defined unambiguously. To overcome theoretical constraints, a set of criteria that are more amenable to a well-founded combination must be found (if the single metric is a must). What we have presented here is the best that can be done without re-examining the objectives.
- **Decision Procedure**: Further development and refinement of the five-step decision-making procedure for generating metrics, and further validation to ensure that the procedure brings out all the critical issues in defining multiple criteria metrics.
- **Industrial Application**: The feasibility of using rigorous decision techniques for making technical decisions needs to be evaluated in an industrial context. The issue here is that

theoretical soundness is in conflict with simplicity and speed of decision making – and it is not obvious what the 'right' trade-offs are (this is another multiple-criteria decision!).

- **Group Decision-making**: Different partners in a project often have different perceptions of the relative priorities of the criteria. Investigation is needed to understand how different stakeholders' opinions (weights as well as operational data) can be aggregated and conflicts resolved for the group to arrive at a decision.

This work was founded on the need to synthesise a hurt metric for avionic equipment controllers but the lessons learned from it are applicable to any product. These include the role of alternatives and criteria, the importance of formal preference modelling and the significance of representation theorems for combining the metric drivers.

11. Acknowledgements

The authors started this work in collaboration with Graham Tanner of Rolls-Royce, but the insights presented here were developed subsequently, under the auspices of the METHOD project, funded by the Engineering and Physical Sciences Research Council of the United Kingdom with grant number GR/J42311).

12. References

[1] Keeney, R.L. and Raiffa, H., "Decisions with Multiple Objectives: Preferences and Value Tradeoffs", Cambridge University Press, 1993.

[2] Vincke, P., "Multicriteria Decision-Aid", John Wiley, 1992.

[3] Saaty, T.L., "Multicriteria Decision Making: The Analytic Hierarchy Process", RWS Publications, Pittsburgh, 1996.

[4] Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A., "Foundations of Measurement: Additive and Polynomial Representations", Volume 1, Academic Press, 1971.

[5] Roberts, F.S., "Measurement Theory with Applications to Decision Making, Utility, and the Social Sciences", Addison-Wesley, 1979.

[6] Prasad, D., "Dependable Systems Integration using Decision Analysis", D.Phil. dissertation in preparation, Department of Computer Science, University of York, U.K.

[7] Ministry of Defence, 00-56, "Safety Management Requirements for Defence Systems" (Part 1: Requirements), Issue 2, December 1996.

[8] "Expert Choice Decision Support Software", Version 9.0, Expert Choice Inc., Pittsburgh, 1995.