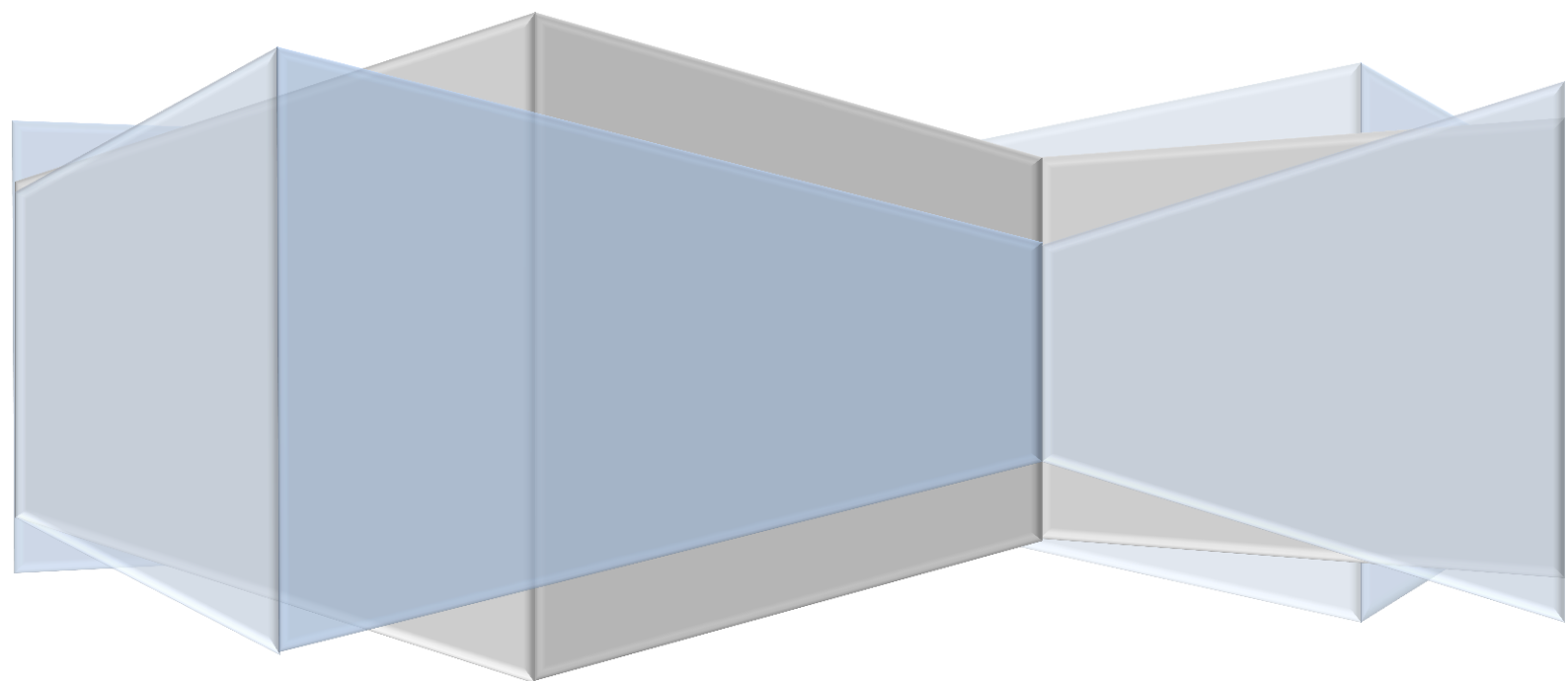


Uncertainty in Assessing the Condition of Critical Infrastructure: The Case of the Forth Road Bridge

John Quigley, University of Strathclyde



Note to the Reader

This case study was prepared for *The Foundations of Risk*, a course offered at the Canada School of Public Service in the winter of 2011. The Foundations of Risk was created by Calvin Burns (University of Strathclyde, UK), John Quigley (University of Strathclyde, UK) and Kevin Quigley (Dalhousie University, Canada).

The development of the case study was led by John Quigley. The case should not be cited without his permission.

Special thanks to Dr Lesley Walls for her invaluable contribution to this case study.

Financial support from the Canada School of Public Service to conduct this work is gratefully acknowledged. The views expressed in this report are not necessarily those of the Canada School of Public Service or of the Government of Canada.

Contact information:

Professor John Quigley
Director of Research
Department of Management Science
University of Strathclyde
40 George Street
Glasgow, G1 1QE

Table of Contents

Learning Aims and Objectives	3
The Forth Road Bridge – Background	3
Statement of the Decision Problem	5
Methodology Used to Build Model	5
Overview of the Model-Building Process	5
Reflections on the Model-Building Process	7
Model Construction	8
Model Structuring: Stages 1 and 2	8
Elicitation of Subjective Probabilities – Stage 3	10
Summarizing the Expert Judgement	13
Recommendation	18
Discussion Questions/Exercises	18
Suggested Answers	19
Bibliography	20

Learning Aims and Objectives

1. To appreciate the role of modelling to support decision-making under uncertainty
2. To reflect on the meaning of probabilistic expert judgement
3. To reflect on the process of elicitation of expert judgement
4. To reflect on the management of multiple experts' judgements
5. To reflect on the role of appropriate summaries in describing uncertainty

The Forth Road Bridge – Background

The Forth Road Bridge is an aging and vital logistical hub in Scotland, providing a key connection between the east coast and the capital city of Edinburgh. The bridge provides crossing over the Firth of Forth for more than 24 million vehicles per year. The bridge first opened in 1964, at which time it was the fourth longest suspension bridge in the world (the longest outside the United States) with a main span of 1006 metres. In total, the structure is over 2.5 km long.



*Image used by permission of 'chiefmoamba' under Creative Commons.
Original image at www.flickr.com/photos/chiefmoamba/75390685/*

Figure 1: Forth Road Bridge

The bridge comprises two main suspension cables, each consisting of almost 12,000 wires. The bridge has four anchorage chambers, two on each shore. Anchorages are concrete-filled tunnels bored into the rock on either shore where the bridge's main suspension cables are attached to the ground.



Figure 2: A suspension cable for Forth Road Bridge entering one of four anchorage chambers

The main cable wires splay out in the anchorage chambers and loop around strand shoes, which are in turn bolted to the face of the concrete tunnels.



Figure 3: A suspension cable splaying as it enters anchorage chamber for Forth Road Bridge

There are 19 crosshead slabs per anchorage chamber and 6 tendons per slab. This is illustrated in Figure 4. The tendons are wedged at the bottom of a rock chamber. The concrete in the tunnel itself is not strong enough to withstand the forces from the cables and was strengthened using pre-tensioned galvanized, high-tensile steel wire strands. Although innovative, it can be vulnerable to corrosion and deterioration in a saline environment, such as is found at Forth.



Figure 4: Inside anchorage for Forth Road Bridge showing the 19 slabs each with 6 tendons bolted to the face of the concrete tunnels

There is guidance from the UK Department of Transport for inspecting post tensioning in bridges as it is acknowledged that there can be problems with this type of construction. The guidelines refer mainly to the difficulties in establishing the condition. These difficulties are exacerbated in a tunnel.

Various problems have been highlighted particularly in relation to early depletion of the bridge. This requires investigation; however, the anchorages' unique design makes this an extremely difficult task.

Statement of the Decision Problem

The current safety of the bridge is not in question. This is a proactive investigation to ensure long-term structural integrity of the anchorages.

There are two tests being considered to assess the condition of the bridge. The first is a direct pull-off test, where the slabs are removed and the tendons, which are behind the slabs, are pulled to assess resistance. The second involves an external excavation of the surrounding area to expose the top row of tendons and inspect them.

The key criterion being used to identify which of the two tests to conduct will be based on which one is most informative with regard to assessing the state of the grout that holds the tendons. The state of the grout cannot be directly observed but can be inferred from test results. Neither test is able to assess the state of the grout perfectly.

The test results will be used to inform recommendations for maintaining the bridge for the remainder of its life. The Scottish parliament has ultimate responsibility for the bridge, but the Forth Estuary Transport Authority (FETA) has responsibility for maintaining the bridge in a safe, efficient and cost-effective manner, while minimizing disruption to traffic.

Methodology Used to Build Model

Overview of the Model-Building Process

A model was required to support the decision of which test to conduct. The model building had three main stages, as illustrated in Figure 5. The first stage involved two facilitators working with a

meta-expert. The meta-expert had substantial expertise both in the breadth and depth of the problem area, while the facilitators had expertise in model construction. The aim of the first stage was to identify the key variables relevant to support the decision and identify the structure of the model, i.e. how the variables interacted with each other.

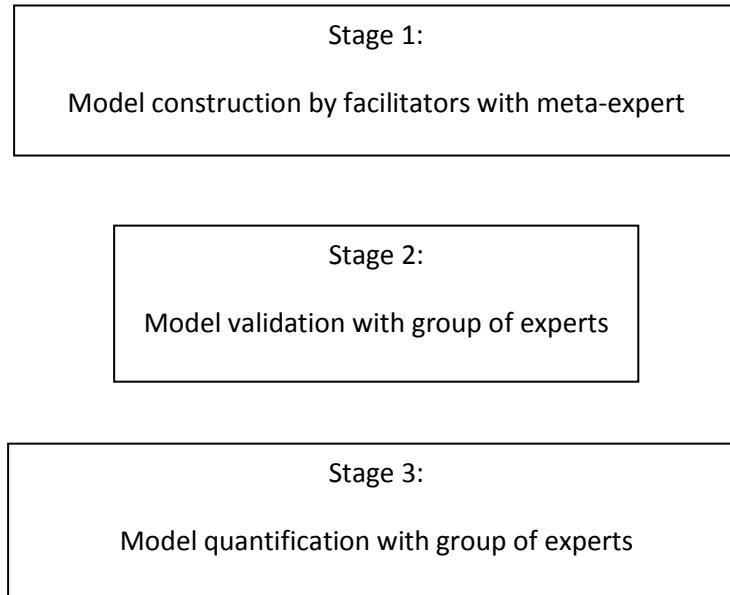


Figure 5: Model-building process

Once the variables were identified and a structure of the model constructed by the meta-expert, the structure was presented to a group of experts with a range of experiences. Views were considered and debated until agreement was obtained.

The third stage of the process involved four experts, all required to quantify their beliefs about the values of the variables in the model.

Expert judgement, like all data collection, can be subject to bias, and the role of the facilitator is to minimize the impact of the bias on the data acquired. For the third stage, the Stanford Research Institute (SRI) suggests a general model for eliciting expert judgement that comprises the following steps relevant for this problem: motivate, structure, condition, encode, verify and aggregate. Each step is designed to minimize the impact of a specific type of bias. The following is a brief summary of each step.

Motivating an expert can be achieved by explaining the elicitation process and how the results will be used. This step of the process can be used to enhance the comprehension of an expert with regard to the process, encourage an expert to provide an accurate assessment of his/her understanding and determine potential biases of an expert through conversation. Examples of motivational biases are: “*expert bias*” where an expert becomes overconfident merely because s/he has been given the title “expert” and “*management bias*” where an expert provides goals espoused by management as opposed to experienced judgement.

Structuring refers to defining the specific event being considered to ensure there is no ambiguity in the questions posed. This step has a second aim, which is to explore how an expert thinks about quantity for which s/he is providing judgement and hence detect *cognitive biases* which can exist due to simplifying the complex task of assigning probabilities (e.g. an expert assigns probabilities for several possible outcomes that when added together sum to greater than 1). Afterwards, disguising the variable through disaggregation into more elementary variables is recommended to combat motivational biases and to reduce cognitive biases.

Judgements can be *conditioned*, which means surfacing all relevant data with the expert to stimulate thinking. This phase can be used to determine biases such as anchoring and availability. “Anchoring” is the assessment of probabilities for several events relative to an initial assessment. For example, an expert starts with the event most likely to occur, assigns a probability value and makes all other probability values relative to this one. These adjustments are typically insufficient. “Availability” is the assignment of higher values than appropriate to events that are more memorable. These two biases result in poor probability assessments because they are indications that an expert is not considering the novelty of the events to be assessed because s/he is relying too heavily on historical data and is ineffectively summarizing expected frequencies. To address these biases, the facilitator should discuss relevant information with the experts.

Encoding refers to an expert providing numerical expression that reflects his/her uncertainty regarding an outcome.

The aim of the fifth step is to *verify* that an expert has provided a reflection of his/her true beliefs. If problems are encountered then the previous steps are to be repeated.

The *aggregation* step refers to aggregating subjective probability distributions from various experts assessing the same situation to obtain one subjective distribution. There are two different approaches to this problem. The first uses mathematical aggregation so that the experts do not influence each other’s decisions or subjective probabilities. A second approach allows experts to share their judgements and re-assess their distributions.

Reflections on the Model-Building Process

The degree of structure inherent in an elicitation process increases as the process is implemented. For example, at the initial phase variables will not have been identified let alone defined, but by the end of the elicitation process implementation we shall have precise definitions of the states of all variables and have identified all conditioning scenarios that require subjective probability distributions.

Typically, the facilitator should possess an understanding of the problem being modelled, albeit not to the same extent as the experts. Moreover, the modelling must have some purpose or goal, of which the facilitator will be cognizant. As such, a fully unstructured approach to elicitation would not be sensible. However, the argument in favour of minimal levels of structure at the start of the elicitation process is that it permits the experts to conceptualize the variables and inform the associated definitions in a way that is meaningful to their experience and reflects their beliefs.

A set of questions to be used as guidelines for facilitators in a semi-structured interview seems appropriate. A semi-structured interview is preferred to a questionnaire or other alternatives because

the set of procedures being developed should be applicable as a guide for an assessor for any project. By employing a semi-structured format, it is acknowledged that all relevant judgements from engineers cannot be anticipated in advance and so, for any interview, flexibility to elicit judgements beyond a strict list of questions is required.

Ultimately, a model is constructed that represents the holistic view of the problem and supports a shared understanding among all experts. There is, therefore, a need to impose some structure on the elicitation process. It is not uncommon, though, to iterate between the problem structuring and instantiation phases as the requirement of quantification can expose ambiguity in definitions and relationships that were not identified during the more qualitative phase of the process.

Model Construction

Model Structuring: Stages 1 and 2

Stage 1 concerns the initial problem-structuring stage, which was conducted with the meta-expert, the Technical Director of the consulting civil engineering company contracted to specify the test design. He has 20 years of experience working with the bridge. A series of eight semi-structured interviews, each lasting two to three hours, were conducted to structure the model qualitatively. There were two facilitators, usually together—one taking detailed notes of the discussion and the other provoking discussion.

The process was new to the experts and initially there was misunderstanding about how their judgements were being represented. The reasoning of the experts was recorded diagrammatically, where variables were represented by oval nodes and arcs linking nodes represented direct causal influences, with arrows showing the direction of the causal relationship. Figure 6 provides a simple example, where one variable might represent the True State of the grout (which can be one of many different states) and this will influence the outcomes of the test.

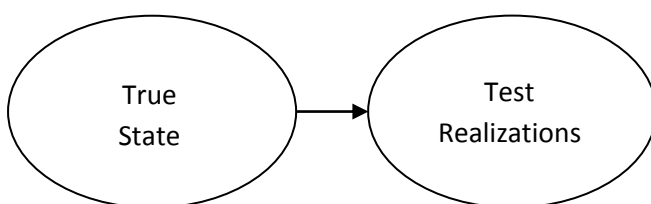


Figure 6: Diagrammatical representation of causal relationship between true state of grout and test realizations

Stage 2 began once we had agreed on a model structure with the meta-expert, where we reviewed it with another six experts. This included the Chief Engineer & Bridgmaster, who is a qualified Civil Engineer, as well as other relevant engineers from both the bridge operations and the contracted civil engineering consultancy. The general methodology was explained and discussed. A facilitated session followed to present the qualitative model and to gather critical independent expert assessments of the logic. All experts accepted the major structures and minor revisions were made to the detail of selected nodes to arrive at the final iteration of the model.

Figure 7 illustrates a simplified version of the agreed qualitative structure of the model. The two nodes on the right-hand side of the network represent the planned tests. The remainder of the nodes include variables that represent the state of different characteristics of the strands and the anchorages that will not be observed directly on test but can be inferred.

From Figure 7 we can see that the experts agree that the condition of the grout on a tendon will affect the results of the direct pull-off test (DPT). Simply, we can consider the state of the grout to be either fully effective or not fully effective, and the results from the DPT are in one of four classifications: fully elastic, juddering, no movement and total failure. These classes describe the experience of pulling a tendon that has been exposed. The model suggests the likelihood that a test outcome will be different for a tendon whose grout is in a fully effective state than for one whose grout was not.

The experts believe that the state of the grout will have an indirect impact on the test outcome from the external excavation test. These test results will be classed into one of three categories: OK, evidence of surface corrosion and no surface corrosion but evidence of pitting. The state of the grout for a tendon will influence the chance of pitting as well as surface corrosion. The presence of pitting will influence the chance of wires being cracked, and cracked wires influence the chance of broken wires being present. From Figure 6, we see that the experts agreed that the presence of surface corrosion, pitting, cracked wires and broken wires all had a direct effect on the test outcomes.

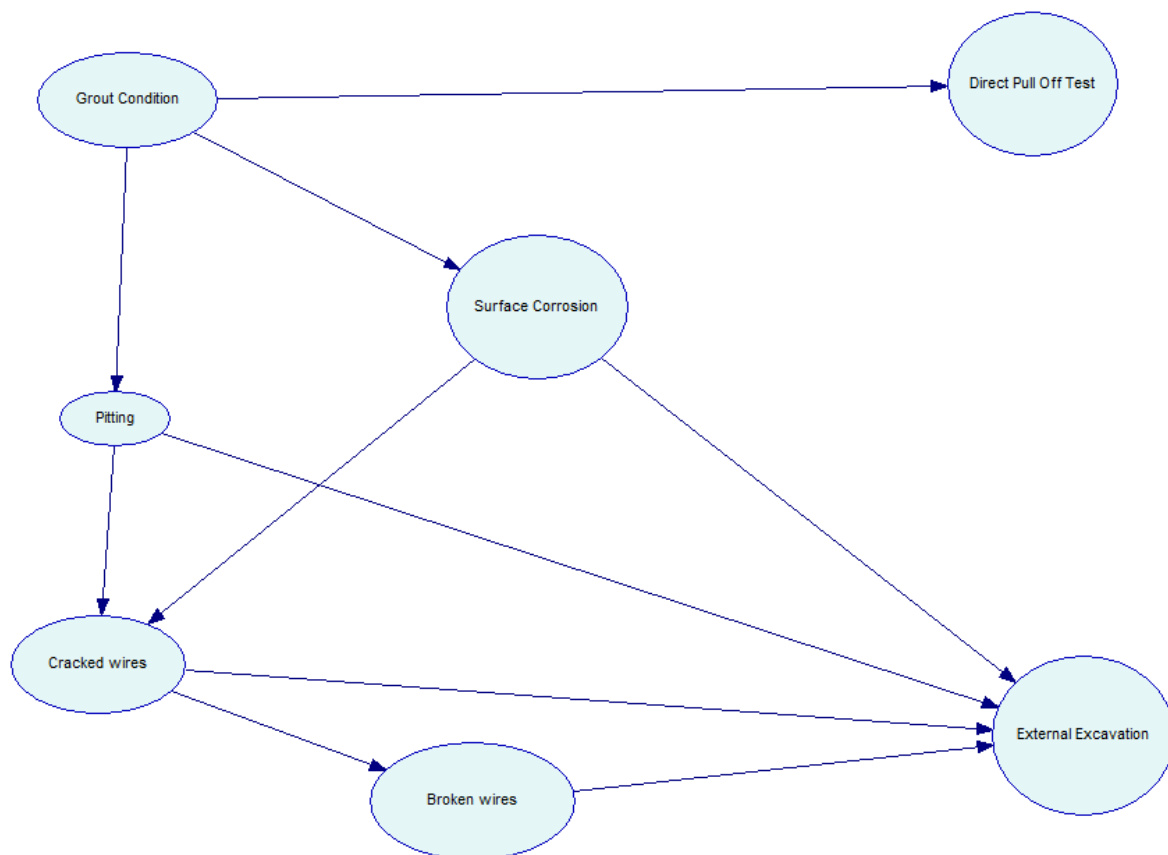


Figure 7: Agreed qualitative structure of the model

Elicitation of Subjective Probabilities – Stage 3

The third stage of the model building concerned quantification. This stage involved multiple engineers, each individually assessing their beliefs and expressing them with probabilities. The experts had been briefed on probability elicitation requirements in a face-to-face meeting. After which they were each provided with a spreadsheet containing probability tables, which they were required to complete.

For each variable in the model there are multiple states. A probability distribution was required from each expert for the proportion of tendons that will be in each state subject to the conditioning scenario. The probability elicitation is extensive and thus demanding on the experts.

Table 1 is a summary of the subjective probabilities provided by each of the four experts in assessing their belief on the proportion of tendons that will be in a fully effective state. Associated with each assessment the expert provided an upper and lower bound on the proportion, such that they did not believe the true proportion of fully effective tendons would be outside the stated upper and lower bounds.

Table 1: Experts' subjective probabilities with bounds on the proportion of tendons in a fully effective state

	Fully effective	Lower bound	Upper bound
Expert 1	0.56	0.45	0.8
Expert 2	0.32	0.2	0.5
Expert 3	0.32	0.2	0.5
Expert 4	0.09	0	0.15

As such, Expert 1 has provided a 'best' assessment that 56% of the tendons are in a fully effective state, while Expert 4's 'best' assessment is that only 9% of the tendons are in a fully effective state. Each expert has provided upper and lower bounds on these estimates reflecting their degree of uncertainty.

Each expert then assessed the *conditional* probability of realizing the different outcomes from the DPT. These are *conditional* probabilities because they are conditioned on different scenarios. Table 2 provides the eight different conditional probability distributions, two from each expert, conditioned on whether the test was being applied to a tendon whose grout was in a fully effective state or not.

Combining the data from Table 1 and Table 2 we have the following. Expert 4's assessment is that 9% (taken from Table 1) of the tendons are Fully Effective of which 23% will have a Fully Effective test outcome (taken from Table 2). As such, Expert 4 thinks that 2% (0.09×0.23) of the tendons will be both in a Fully Effective state and produce a test result of Full Elastic.

Table 2: Conditional probabilities provided by each expert for direct pull-off test

Expert Test outcome/ grout state	1		2		3		4	
	FE	NFE	FE	NFE	FE	NFE	FE	NFE
Fully elastic	0	0.37	0.77	0.67	0.85	0.76	0.23	0.48
Juddering	0.98	0.43	0.01	0.11	0.1	0.15	0.58	0.24
No movement	0	0	0.17	0.05	0.01	0.02	0.09	0.04
Total failure	0.02	0.2	0.04	0.17	0.04	0.07	0.1	0.24

FE = Fully Effective; NFE = Not Fully Effective

To assess the conditional probabilities of the external excavation test (EET), it was necessary to elicit the conditional probabilities for all intermediate nodes first. The resulting eight conditional probabilities for the EET are summarized in Table 3.

Table 3: Conditional probabilities provided by each expert for external excavation test

Expert Test outcome/ grout state	1		2		3		4	
	FE	NFE	FE	NFE	FE	NFE	FE	NFE
OK	0.86	0.16	0.1	0.01	0.67	0.23	0.34	0.06
Surface corrosion present	0.12	0.32	0.6	0.68	0.2	0.45	0.61	0.87
No surface corrosion but pitting	0.02	0.52	0.3	0.31	0.13	0.32	0.05	0.07

FE = Fully Effective; NFE = Not Fully Effective

We see from Table 3 that Expert 4 thinks 34% of all fully effective tendons will have a test result of OK. As such, Expert 4 thinks that 3% (0.09×0.34) of tendons are fully effective and will result in a test result of OK.

In Table 4 (a to d) we have summaries of the probability distributions for both state of the grout and the associated test outcome obtained using the data from each expert. Table 4a shows the probabilities from Expert 1. We see that the probability that a fully effective tendon is selected to be tested and the resulting test result of fully elastic is 0, i.e. Expert 1 does not believe this can happen. Through summing the probabilities down the column Fully Effective we obtain the probability that a fully effective tendon is selected. Likewise through summing the probabilities of Fully Effective and Not Fully Effective for any row we obtain the probability of a test outcome regardless of the state of the tendon being tested, which are summarized in the column Total. For example, the probability of a Juddering test outcome is 0.7380, while the probability of a Total Failure is 0.0992. Finally, for each test outcome we have calculated the proportion that is Fully Effective. Viewing the results in the last column, consider the following, from Table 4a: Expert 1 expects that in a sample of 10,000 tendons we would observe 7380 testing Juddering and of these 7380, we expect 5488 would be Fully Effective, i.e. 74.36% of the 7380.

Table 4a: Probability distribution of the state of the grout and the DPT outcome for Expert 1

Test Outcome/ Grout State	Fully Effective	Not Fully Effective	Total	Proportion of outcomes Fully Effective
Fully Elastic	0	0.1628	0.1628	0
Juddering	0.5488	0.1892	0.7380	0.7436
No Movement	0	0	0	
Total Failure	0.0112	0.088	0.0992	0.1129

The final column of this table, i.e. Proportion of outcomes Fully Effective, is quite important in assessing the efficacy of the test. We will only be able to observe the test result not the actual state of the grout. As such, if we receive a test result of Fully Elastic, according to Expert 1 it must be Not Fully Effective, but if the test outcome is Juddering then there would be a 54.88% chance the tendon was Fully Effective. The better the test, the closer this value will be to either 0 or 1, as it is better at discriminating between the different states of grout.

Table 4b: Probability distribution of the state of the grout and the DPT outcome for Expert 2

Test Outcome/ Grout State	Fully Effective	Not Fully Effective	Total	Proportion of outcomes Fully Effective
Fully Elastic	0.2464	0.4556	0.7020	0.3510
Juddering	0.0032	0.0748	0.0780	0.0410
No Movement	0.0544	0.0340	0.0884	0.6154
Total Failure	0.0128	0.1156	0.1284	0.0997

Table 4c: Probability distribution of the state of the grout and the DPT outcome for Expert 3

Test Outcome/ Grout State	Fully Effective	Not Fully Effective	Total	Proportion of outcomes Fully Effective
Fully Elastic	0.2720	0.5168	0.7888	0.3448
Juddering	0.0320	0.1020	0.1340	0.2388
No Movement	0.0032	0.0136	0.0168	0.1905
Total Failure	0.0128	0.0476	0.0604	0.2119

Table 4d: Probability distribution of the state of the grout and the DPT outcome for Expert 4

Test Outcome/ Grout State	Fully Effective	Not Fully Effective	Total	Proportion of outcomes Fully Effective
Fully Elastic	0.0207	0.4368	0.4575	0.0452
Juddering	0.0522	0.2184	0.2706	0.1929
No Movement	0.0081	0.0364	0.0445	0.1820
Total Failure	0.0090	0.2184	0.2274	0.0396

Similar calculations were conducted with the conditional probabilities for the External Excavation Test (EET) outcomes. These are summarized in Table 5 (a-d) for each expert.

Table 5a: Probability distribution of the state of the grout and the EET outcome for Expert 1

Test Outcome/Grout State	Fully Effective	Not Fully Effective	Total	Proportion of outcomes Fully Effective
OK	0.4816	0.0704	0.5520	0.8725
Surface Corrosion present	0.0672	0.1408	0.2080	0.3231
No Surface Corrosion but pitting	0.0112	0.2288	0.2400	0.0467

Table 5b: Probability distribution of the state of the grout and the EET outcome for Expert 2

Test Outcome/Grout State	Fully Effective	Not Fully Effective	Total	Proportion of outcomes Fully Effective
OK	0.0320	0.0068	0.0388	0.8247
Surface Corrosion present	0.1920	0.4624	0.6544	0.2934
No Surface Corrosion but pitting	0.0960	0.2108	0.3068	0.3129

Table 5c: Probability distribution of the state of the grout and the EET outcome for Expert 3

Test Outcome/Grout State	Fully Effective	Not Fully Effective	Total	Proportion of outcomes Fully Effective
OK	0.2144	0.1564	0.3708	0.5782
Surface Corrosion present	0.0640	0.3060	0.3700	0.1730
No Surface Corrosion but pitting	0.0416	0.2176	0.2592	0.1605

Table 5d: Probability distribution of the state of the grout and the EET outcome for Expert 4

Test Outcome/Grout State	Fully Effective	Not Fully Effective	Total	Proportion of outcomes Fully Effective
OK	0.0306	0.0546	0.0852	0.3592
Surface Corrosion present	0.0549	0.7917	0.8466	0.0648
No Surface Corrosion but pitting	0.0045	0.0637	0.0682	0.0660

Summarizing the Expert Judgement

Thus far we have a table of probabilities for each expert for each test. We need to summarize these data with the purpose of assessing which test will discriminate between the states of the grout better than the other.

Consider Expert 4 who has assessed that 9% of the tendons are fully effective. From Table 4d we see that the DPT will result in an updated estimate of between 4% (i.e. 0.0396) and 19%, while from Table 5a we see that the EET will result in an updated estimate of between 6% and 36%. From such a summary it would appear that EET has a greater potential to shift the opinion of the expert and as such be the better test. However, for the EET there is a probability of 0.0852 that the updated estimate will be 0.3592 and therefore a probability of 0.9148 that the updated probability will be 0.06. As such, comparing the tests on the range of updated assessments might not be the most appropriate.

The measure used to assess the variability in updating was the standard deviation, which is a commonly used statistic for measuring dispersion. To determine the standard deviation we first calculate the difference between the initial estimate (for Expert 4 this would be 0.09) and each of the possible updated estimates. Each difference is then squared and a weighted average of the squared differences is calculated with the probabilities being used as weights. From this we obtain an average squared difference. Finally, the square root is taken of the average squared difference to express the summary statistic in the original units of measures.

Calculation of the standard deviation of the DPT for Expert 4 is illustrated in the following.

$$\begin{aligned}
 st\ dev &= \sqrt{(0.0452 - 0.09)^2 \times 0.4575 + (0.1929 - 0.09)^2 \times 0.2706} \\
 &\quad + (0.1820 - 0.09)^2 \times 0.0445 + (0.0396 - 0.09)^2 \times 0.2274 \\
 &= 0.0688
 \end{aligned}$$

Using this as the summary measure of efficacy for the tests we can compare the tests for each expert. Figure 8 is a decision tree representing the possible consequences from the different choice of test for Expert 4. The squared difference between the initial assessment by the expert and the updated assessment is recorded at the end of each branch. The standard deviation from each test is summarized in each circle.

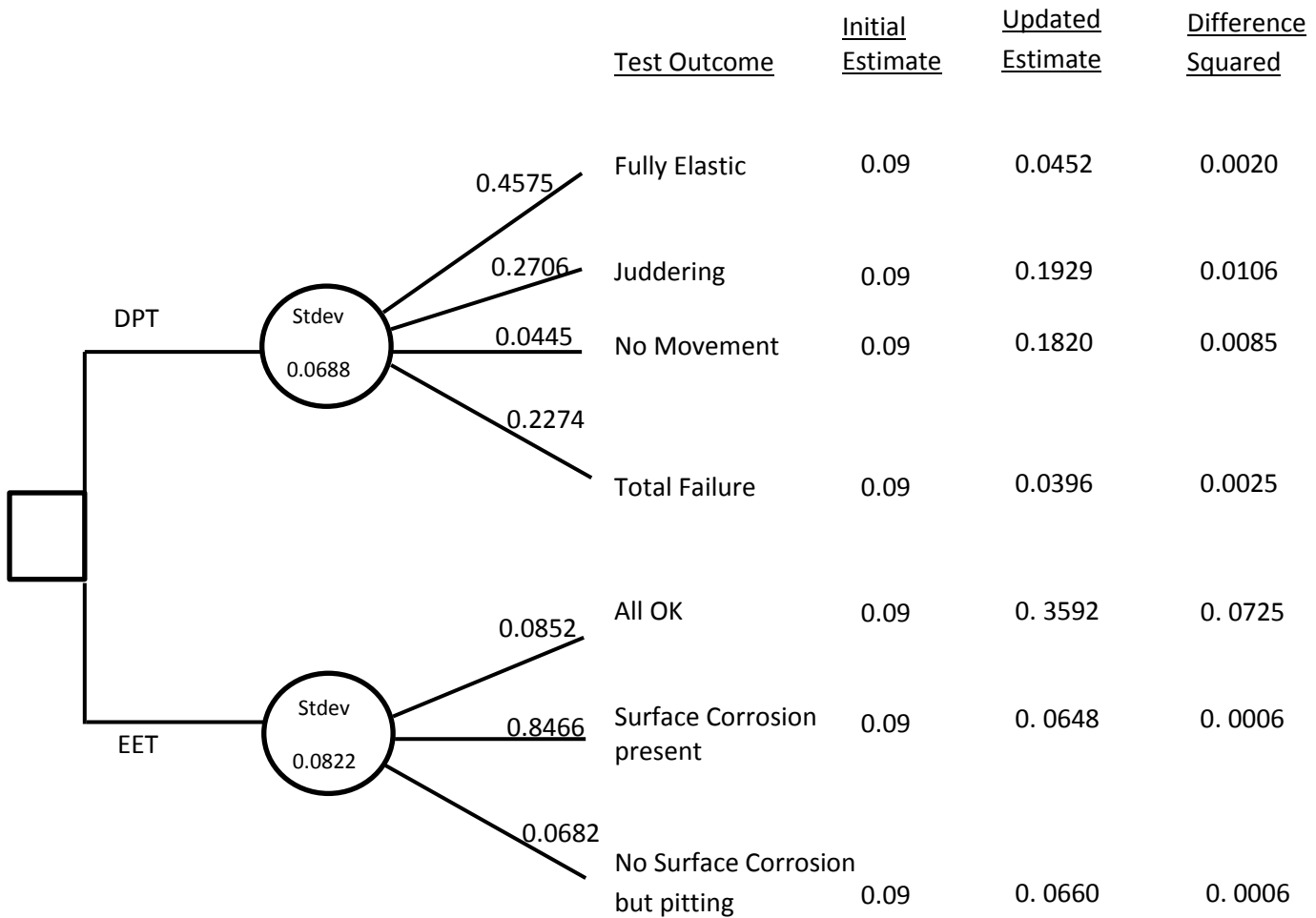


Figure 8: Decision tree for Expert 4 showing EET to be the optimal test to assess the proportion of tendons that are fully effective

Similar decision trees can be produced for each of the other three experts. The resulting data and standard deviations used to inform the decision are summarized in the following tables. In Table 6 (a-d) are the calculations used to derive the standard deviation for the DPT using the data from Experts 1 to 4.

Table 6a: Expert 1 Evaluation of direct pull-off test

Direct Pull-Off Test Results	Probability	Updated	Difference ²
Fully Elastic	0.1628	0	0.3136
Juddering	0.7380	0.7436	0.0337
No Movement	0		0.3136
Total Failure	0.0992	0.1129	0.1999
Standard Deviation			0.3095

Table 6b: Expert 2 Evaluation of direct pull-off test

Direct Pull-Off Test Results	Probability	Updated	Difference ²
Fully Elastic	0.7020	0.3510	0.0010

Juddering	0.0780	0.0410	0.0778
No Movement	0.0884	0.6154	0.0873
Total Failure	0.1284	0.0997	0.0485
Standard Deviation			0.1438

Table 6c: Expert 3 Evaluation of direct pull-off test

Direct Pull-Off Test Results	Probability	Updated	Difference ²
Fully Elastic	0.7888	0.3448	0.0006
Juddering	0.1340	0.2388	0.0066
No Movement	0.0168	0.1905	0.0168
Total Failure	0.0604	0.2119	0.0117
Standard Deviation			0.0485

Table 6d: Expert 4 Evaluation of direct pull-off test

Direct Pull-Off Test Results	Probability	Updated	Difference ²
Fully Elastic	0.4575	0.0452	0.0020
Juddering	0.2706	0.1929	0.0106
No Movement	0.0445	0.1820	0.0085
Total Failure	0.2274	0.0396	0.0025
Standard Deviation			0.0688

We see that Expert 1 has the largest standard deviation based on the outcomes of the DPT while Expert 3 has the smallest.

Tables 7 (a-d) present the probabilities and standard deviations for the EET based on the data provided by the four experts.

Table 7a: Expert 1 Evaluation of EET

OSI	Probability	Updated	Difference ²
All OK	0.5520	0.8725	0.0976
Surface Corrosion present	0.2080	0.3231	0.0561
No Surface Corrosion but pitting	0.2400	0.0467	0.2635
Standard Deviation			0.3589

Table 7b: Expert 2 Evaluation of EET

OSI	Probability	Updated	Difference ²
All OK	0.0388	0.8247	0.2548
Surface Corrosion present	0.6544	0.2934	0.0007
No Surface Corrosion but pitting	0.3068	0.3129	0.0001
Standard Deviation			0.1018

Table 7c: Expert 3 Evaluation of EET

OSI	Probability	Updated	Difference ²
All OK	0.3708	0.5782	0.0667
Surface Corrosion present	0.3700	0.1730	0.0216
No Surface Corrosion but pitting	0.2592	0.1605	0.0254
Standard Deviation			0.1983

Table 7d: Expert 4 Evaluation of EET

OSI	Probability	Updated	Difference ²
All OK	0.0852	0.3592	0.0724
Surface Corrosion present	0.8466	0.0648	0.0006
No Surface Corrosion but pitting	0.0682	0.0660	0.0006
Standard Deviation			0.0821

Inspection of the tables shows that only Expert 2 would recommend DPT, while all others would recommend the EET.

Recommendation

The standard deviations of the updated probabilities are summarized in Table 8 for both tests. This suggests that greater the standard deviation, the greater the movement from the initial estimate and the more informative the test. Consider a test, after which we are guaranteed to have the same uncertainties as we have before the test; such a test would be completely uninformative and this would be reflected in a standard deviation of 0.

Table 8: Standard deviation of updated probability of item being tested having grout condition of fully effective

Expert	DPT	EET
1	0.3095	0.3589
2	0.1438	0.1018
3	0.0485	0.1983
4	0.0688	0.0821

From viewing Table 8 we see that only Expert 2 thinks that the DPT would be more informative than the EET. As such, the recommendation is for the EET. Expert 3 shows the greatest difference with the standard deviation for the EET being four times the value of the standard deviation for the DPT.

Discussion Questions/Exercises

1. Investigate the difference between aleatory and epistemic uncertainty. Describe each in the context of this case. Explain why it is important to distinguish between these two types of uncertainty.
2. Why is the elicitation of probabilities a socio-technical exercise?
3. The range of subjective probabilities presented in Table 1 is disjointed in so far as the intervals provided by each expert do not overlap with each other, e.g. Expert 1 provides a range of 0.45 to 0.8 while Expert 2 provides a range of 0.2 to 0.5. To what extent can we say that all experts are correct in their assessment?
4. Discuss the implications of averaging all the input assessments before working through the calculations to obtain a recommendation. Would this be more sensible?
5. The upper and lower bounds of the initial assessment in Table 1 were not used. How could they be used in this study?
6. In section 2, the responsibilities of FETA were summarized. Reflecting on those responsibilities, was narrowly defining the criterion for this study appropriate? How might such a multi-criteria decision problem be operationalized?
7. How should the results of this study be communicated to the Minister of Transport or the general public? What challenges does this present?
8. Could this decision problem have been adequately supported using an entirely qualitative method?

Suggested Answers

1. Aleatory uncertainty corresponds to inherent randomness in the process generating observations and epistemic uncertainty corresponds to state of knowledge of an expert. In the context of testing, even if we knew the exact proportion of items that would generate certain test outcomes there would still be variability from one test outcome to another. This is aleatory uncertainty, which we cannot control but can reduce with more samples. There are some uncertainties within the model that cannot be reduced, as the actual states of the grout are not observable and as such we may not be able to test each expert's link between the grout and the test results. It is important to be able to understand how data will be able to distinguish between different expert assessments.
2. Humans are not natural probability assessors; they have natural biases. Converting their experiences into an appropriate form from which a numerical expression can be elicited is not trivial and requires a facilitator to work through the process with the expert, challenging and explaining throughout.
3. A subjective probability is a degree of belief held by an expert; as such there is no requirement that it agrees with any other expert's subjective probability. Often, when assessing the accuracy of an expert's assessment we measure the frequency of events being realized for which an expert has assigned a certain probability, for an expert to be calibrated means for example we would expect that 10% of occasions when a calibrated expert assigns a probability of 10% to an event it is realized and 90% of such events are not realized. However, the frequency of the realizations exceeds the assessments of the experts. When we consider the context of a single tendon then it is either from a situation where the grout is fully effective or not; however, when we consider the collection of tendons then the experts have provided a prediction on the proportion that will be fully effective; these can be compared for accuracy. In sum, we cannot assess whether an expert is right or wrong about a single tendon, but we can when we consider many tendons.
4. One could obtain a different recommendation by averaging the data before the analysis. It is tempting to simplify the analysis to create the *average expert*, but we need to consider that this *average expert* is no one's opinion and as such no expert would stand behind the resulting recommendation.
5. A range of values can be used to conduct a sensitivity analysis. The calculations can be repeated for the upper and lower values to assess whether the recommendations are sensitive to these inputs. If the recommendation changes it may be worth bringing the experts together to assess how the uncertainty can be reduced.
6. Two other relevant criteria are cost of the test and disruption to traffic. The key question becomes how much the decision-making authority would be willing to trade off between the test that is anticipated to be the most informative and the other criteria. Such a process may require the involvement of those affected by the disruption through a *town hall* event. Elicitation of the values a community will place on disruptions, costs and information from a test and their willingness to trade-off between each will also be a socio-technical exercise and non-trivial.
7. Expect a discussion of the different ways in which uncertainty is expressed or ignored in communication, with a reflection on the trade-off between providing information and just

providing data. The implications of simplification are that the recommendations can appear more certain than appropriate.

8. The experts learn as they work through the process. At the start of the model building they are not in a position to assess which test is best. Through identifying the critical features of the problem, organizing them in a coherent manner and measuring their impact on the problem we are able to make a recommendation. Qualitative methodologies lack a facility to measure impact and combine measurement adequately; mathematics is required.

Bibliography

The Forth Road Bridge website contains reports detailing the testing and investigation: <http://www.forthroadbridge.org/home>.

The model presented in this case study is simplified for the purposes of meeting the learning aims and objectives. A paper detailing the model was presented at the 2010 ASRANet Conference in Edinburgh and published in its Proceedings.

Quigley, J. and Walls, L. (2010) "Reconciling experts opinion concerning the value of testing using Bayesian Networks", 5th International ASRANet Conference, Edinburgh.