Description of the Expert elicitation Results for the Flight Crew Error Model

Oswaldo Morales
EWI, TU Delft,
August , 2007

Quantification of the Flight Crew Error Model.

This document reports briefly on the final quantification of the Flight Crew Error Performance Model. The model is presented in Figure 1. A summary of the variables entered in the model is presented in table 1. Marginal distributions for which data is not available have been elicited through structured expert judgment [1]. The (un)conditional rank correlations shown in Figure 1 have been elicited through conditional probabilities of exceedence [2]. In total, five experts have been elicited and the results of the expert judgment exercise will be discussed.

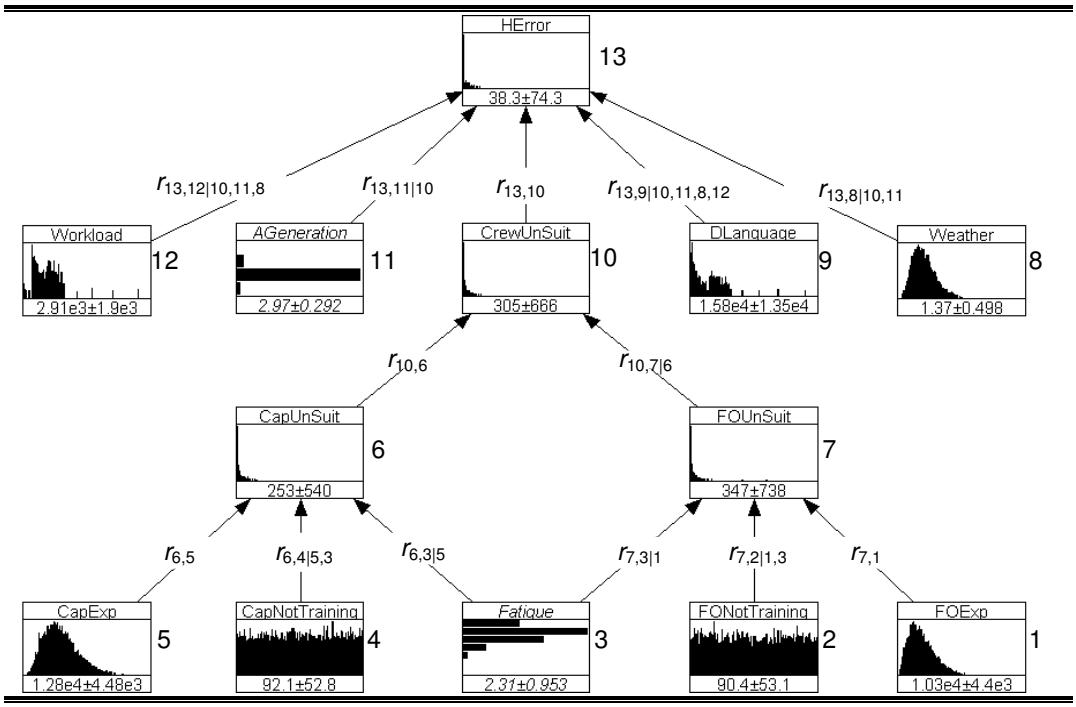


Figure 1. Flight Crew Performance Model.

Name	Description	Source
FOExp	Total number of hours flown since the pilot's license obtaining by first officers.	Data
FONotTraining	Number of days passed since last recurrence training for First Officers.	Data
Fatigue	Stanford Sleepiness Scale. 1signifies "feeling active and vital; wide awake" and 7 stands for "almost in reverie; sleep onset soon; struggle to remain awake".	Data
CapNotTraining	Number of days passed since last recurrence training for Captains.	Data
CapExp	Total number of hours flown since the pilot's license obtaining by captains.	Data
CapUnSuit	Number of Captains failing their proficiency check test per 10,000	Expert Judgment
FOUnSuit	Number of First Officers failing their proficiency check test per 10,000	Expert Judgment
Weather	Rainfall rate (mm/hr) translated into airborne weather radar in a cockpit.	Data
DLanguage	Number of flights in which the pilot and first officer will have a different mother tongue per 100,000	Expert Judgment
CrewUnSuit	Number of Captains or/and First Officers failing their proficiency check test per 10,000	Expert Judgment
AGeneration	Aircraft generation is a scale from 1 to 4 where 4 is the most recent generation of aircrafts.	Data
Workload	Number of times the crew members have to refer to the abnormal/emergency procedures section of the aircraft operation manual during flight per 100,000 flights.	Expert Judgment
HError	Number of errors per 1e6 flights. A distribution is fitted to data provided by DNV on average number of human errors per million flight and its 90% confidence intervals. Brakes not applied correctly per take off rejection with A/C system failure in this example	Data, DNV

Table 1 Flight Crew Performance Model Variables.

THE CLASSICAL MODEL FOR EXPERT JUDGMENT.

The so-called Classical Model [1] for deriving uncertainty distributions over model parameters from expert judgements will be used. The Classical model is a performance based linear pooling or weighted averaging model. The weights are derived from experts' calibration and information performance, as measured on calibration or seed variables. These are variables from the experts' field whose values become known to the experts *post hoc*. Seed variables serve a threefold purpose: (i) to quantify experts' performance as subjective probability assessors, (ii) to enable performance-optimised combinations of expert distributions, and (iii) to evaluate and hopefully validate the combination of expert judgements.

Calibration measures the statistical likelihood that a set of experimental results corresponds, in a statistical sense, with the experts' assessments. In particular, the calibration score is the p-value of a standard Chi-square goodness of fit test. Loosely, the calibration score is the probability that the divergence between the expert's probabilities and the observed values of the seed variables might have arisen by chance. A high score (near one, but bigger than, say, 0.05) means that the expert's probabilities are statistically supported by the set of seed variables. Information represents the degree to which an expert's distribution is concentrated, relative to some user-selected background measure.

"Good expertise" corresponds to good calibration (high statistical likelihood) and high information. Highly informative distributions are desirable only if they are well calibrated. Information should discriminate between more or less equally well calibrated experts. The weights in the classical model are proportional to the product of statistical likelihood and information, and it turns out that this product is indeed dominated by calibration. When a combined expert has been formed, we can also measure the calibration and information of this combined expert. For more detail see [1]. Calculations are performed with the EXCALIBUR software developed at TUDelft.

EXPERT'S PERFORMANCE.

Table 2 below, shows the calibration and information scores for the six experts in this study. The first and second columns give the expert number and id respectively; the third column gives the calibration score. The ratio of highest to lowest score is about 1.30E+04. It will be noted that experts *B* and *D* had a score corresponding to a p-value above 5%. Scores of Experts *E* and *C* are marginal and for expert *A* it is rather low. Calibration scores in the order 0.001 would fail to confer the requisite level of confidence in the results.

The information scores for all items and for calibrations items are shown in columns 4 and 5 respectively. It will be noted that the overall information scores are quite similar, within a factor 2. In this case the expert with the best calibration score (nr 4) also has one of the lowest information scores for the calibration variables which is a recurrent pattern. The last column gives the "un-normlaized weight"; this is the product of columns 3 and 5. If this column were normalized and used to form weighted combinations, experts *A*, *D* and *B* would be influential with (2.25, 32.49 and 64.98 per cent respectively).

Results of scoring experts
Bayesian Updates: no Weights: equal DM Optimization: no
Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relati total	Mean relati realizatioo	Numb real	UnNormalize weight
1	C	0.001547	1.016	0.9689	8	0.001499
2	A	0.02651	0.7119	0.4991	8	0.01323
3	D	0.185	1.317	1.029	8	0.1905
4	B	0.6638	0.95	0.574	8	0.381
5	E	5.115E-005	1.049	1.06	8	5.422E-005

(c) 1989-2005 TU Delft

Table 2. Expert's Performance.

COMBINATIONS SCHEMES.

In this exercise, experts give their uncertainty assessments on calibration variables in the form of 5%, 50%, and 95% quantiles. To combine all experts' assessments into one uncertainty assessment there are three combination schemes. The combined distributions are weighted sums of the individual experts' distributions, with non-negative weights adding to one. Different combination schemes are distinguished by the method according to which the weights are assigned to densities. These schemes are designated "Decision Makers". Two kinds of decision makers are described below.

EQUAL WEIGHT DECISION MAKER.

The equal weight decision maker (EWDM) results by assigning equal weight to each density. In Table 3 the 8th expert is identified as "EWDM". As it may be seen in table3 the EWDM is better calibrated than each expert individually except expert *B*. However information scores derived from the EWDM are poor. They are the lowest amongst 6 experts in both all variables and calibration questions alone. Weights for each expert and for the decision maker are shown in the last three columns. Since the equal weight combination has been chosen in this case, column 8 displays the same weight for all experts except the decision maker, and finally column 9 shows that when normalizing the numbers in column 7 the decision maker would have a weight similar to that of expert *A*.

Results of scoring experts
Bayesian Updates: no Weights: equal DM Optimization: no
Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relati total	Mean relati realizatioo	Numb real	UnNormalize weight	Normaliz.we without DM	Normaliz.we with DM
1	C	0.001547	1.016	0.9689	8	0.001499	0.2	0.002465
2	A	0.02651	0.7119	0.4991	8	0.01323	0.2	0.02175
3	D	0.185	1.317	1.029	8	0.1905	0.2	0.3131
4	B	0.6638	0.95	0.574	8	0.381	0.2	0.6263
5	E	5.115E-005	1.049	1.06	8	5.422E-005	0.2	8.912E-005
6	EWDM	0.2224	0.1046	0.09945	8	0.02212		0.03636

(c) 1989-2005 TU Delft

Table 3. Expert's and EW Decision Maker's Performance.

GLOBAL WEIGHT DECISION MAKER.

This kind of decision maker is from the class of performance based decision makers that are those where the weights are based on the experts' performance on calibration variables. Two performance based decision makers are supported in the software EXCALIBUR. The "global weight" decision maker uses average information over all calibration variables and computes one set of weights for all items. The "item weight" decision maker constructs weights for each item separately, using the experts' information scores for the given item, rather than the average information score.

In this study the global and items weights do not differ, and we focus on the former. The global weight decision maker (Table 4) uses performance based weights which are defined, per expert, by the product of expert's calibration score and his(her) overall information score on calibration variables, and by an optimization procedure[1]. In this case, all experts with a calibration score less than the significance level found by the optimization procedure are outweighed as reflected by the zeros in columns 7, 8 and 9 in Table 4.

In table 4 one can sere that the optimized decision maker gives all weight to expert *B*. Comparing tables 3 and 4 one can see that the calibration score of the GWDM is about 3 times higher than the EWDM and the information score is about 9 times higher over all variables and 5.7 times higher in calibration questions alone.

Table 4.1 presents scores for the GWDM without optimization. In this case, after normalization of the weights of each expert one may observe that, experts *A*, *D*, *B* and the *GWDM_NoO* would be influential with 2.06, 29.66, 59.32 and 8.71 per cent respectively. Though more experts enter the pool, the calibration score of the *GWDM_NoO* is 4.58 times smaller than that of the optimized GWDM. Similarly the information scores in all variables and calibration variables are 2.57 and 1.48 times higher in the optimized decision maker. The GWDM performs better in both information and calibration scores.

[1] The weights used in the classical model are *proper scores*. A score is called *strictly proper* if an expert can achieve his highest expected score by and only by stating his true opinion. The measures for information and calibration described above must be combined in such a way that the result is (in the long run) a strictly proper scoring rule. This requires that the measurement of calibration be combined with "classical significance tests". Briefly, there must be some value $\alpha > 0$, such that if the expert's statistical likelihood drops below α, his/her weight becomes zero. For each value of α it is defined a decision maker dm_α computed as a weighted linear combination of the experts whose calibration score exceeds α. dm_α is scored with respect to calibration and information. The weight which this dm_α would receive if he were added as a "virtual expert" is called the "virtual weight" of dm_α. The value of α for which the virtual weight of dm_α is the greatest is chosen as the cut-off value for determining the unweighted expert.

Results of scoring experts and Relative Information to the DM
Bayesian Updates: no Weights: global DM Optimization: yes
Significance Level: 0.6638 Calibration Power: 1

Nr.	Id	Calibr.	Mean relati total	Mean relati realizatioo	Numb real	UnNormalize weight	Normaliz.we without DM	Normaliz.we with DM	Rel.Inf to total	Rel.Inf to realiz.
1	C	0.001547	1.016	0.9689	8	0	0	0	2.022	1.882
2	A	0.02651	0.7119	0.4991	8	0	0	0	1.962	1.177
3	D	0.185	1.317	1.029	8	0	0	0	2.154	1.114
4	B	0.6638	0.95	0.574	8	0.381	1	0.5	0	0
5	E	5.115E-005	1.049	1.06	8	0	0	0	2.029	1.69
6	GWDM	0.6638	0.95	0.574	8	0.381		0.5	0	0

(c) 1989-2005 TU Delft

Table 4. Expert's and GW Decision Maker's Performance.

Case name : ATCEP Error 5 experts 31 jan 086/3/2008 Version W1.0

Results of scoring experts
Bayesian Updates: no Weights: global DM Optimisation: yes
Significance Level: 0.01566 Calibration Power: 1

Nr.	Id	Calibr.	Mean relati total	Mean relati realizatioo	Numb real	UnNormalize weight	Normaliz.we without DM	Normaliz.we with DM
1	A	0.02651	0.659	0.5962	8	0.01581	0.5021	0.08047
2	B	0.01566	1.083	1.001	8	0.01568	0.4979	0.07981
3	C	0.003315	1.489	1.405	8	0	0	0
4	D	5.857E-007	1.659	1.641	8	0	0	0
5	E	2.309E-005	1.082	1.022	8	0	0	0
6	GWDM	0.6638	0.3468	0.2485	8	0.165		0.8397

(c) 1989-2005 TU Delft

Results of scoring experts
Bayesian Updates: no Weights: global DM Optimisation: no
Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relati total	Mean relati realizatioo	Numb real	UnNormalize weight	Normaliz.we without DM	Normaliz.we with DM
1	C	0.001547	1.016	0.9689	8	0.001499	0.002558	0.002335
2	A	0.02651	0.7119	0.4991	8	0.01323	0.02258	0.02061
3	D	0.185	1.317	1.029	8	0.1905	0.3249	0.2966
4	B	0.6638	0.95	0.574	8	0.381	0.6499	0.5932
5	E	5.115E-005	1.049	1.06	8	5.422E-005	9.249E-005	8.442E-005
6	GWDM_NoO	0.1447	0.3689	0.3868	8	0.05599		0.08718

(c) 1989-2005 TU Delft

Table 4.1 Expert's and GW Decision Maker's (without optimization) Performance.

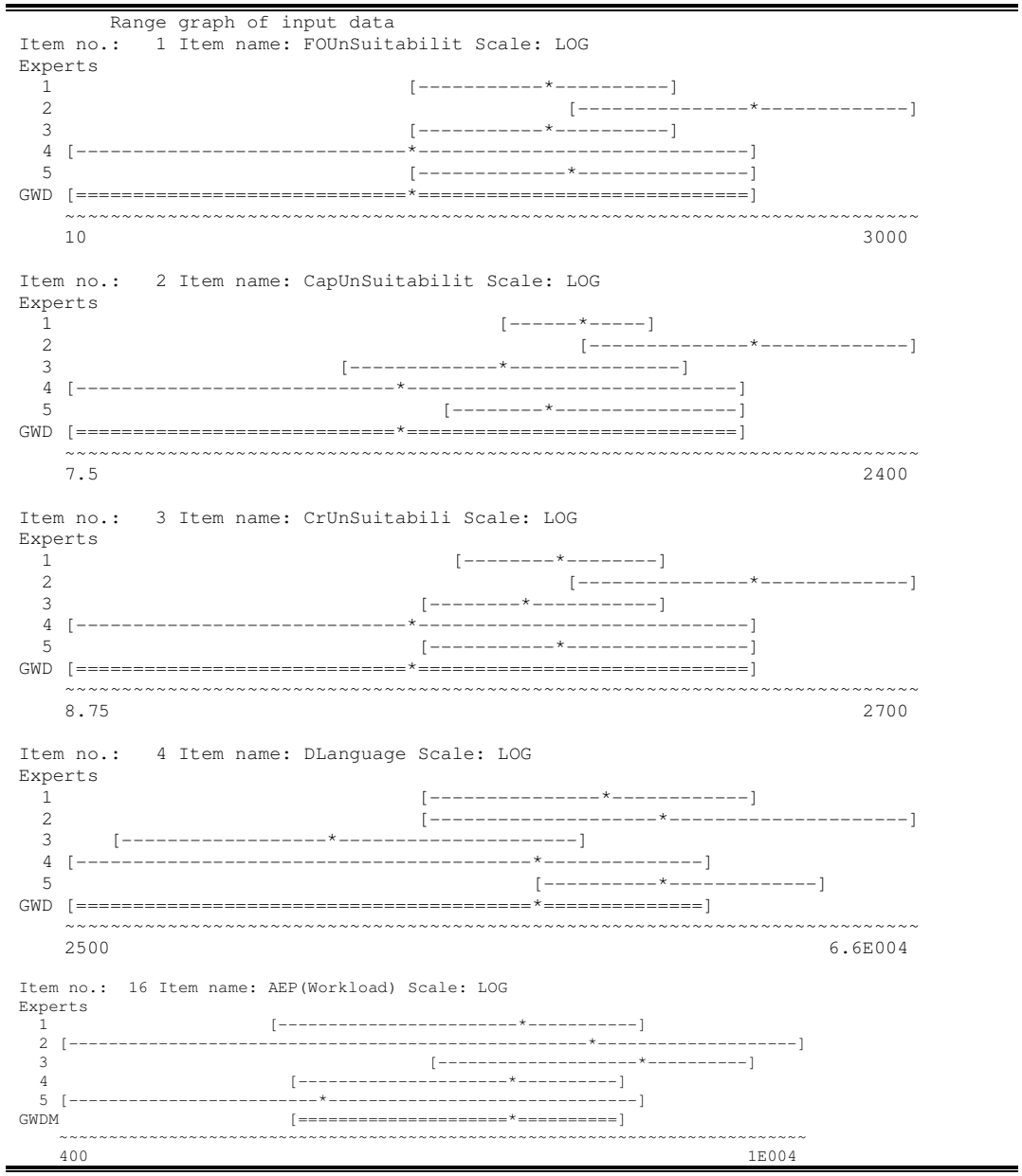
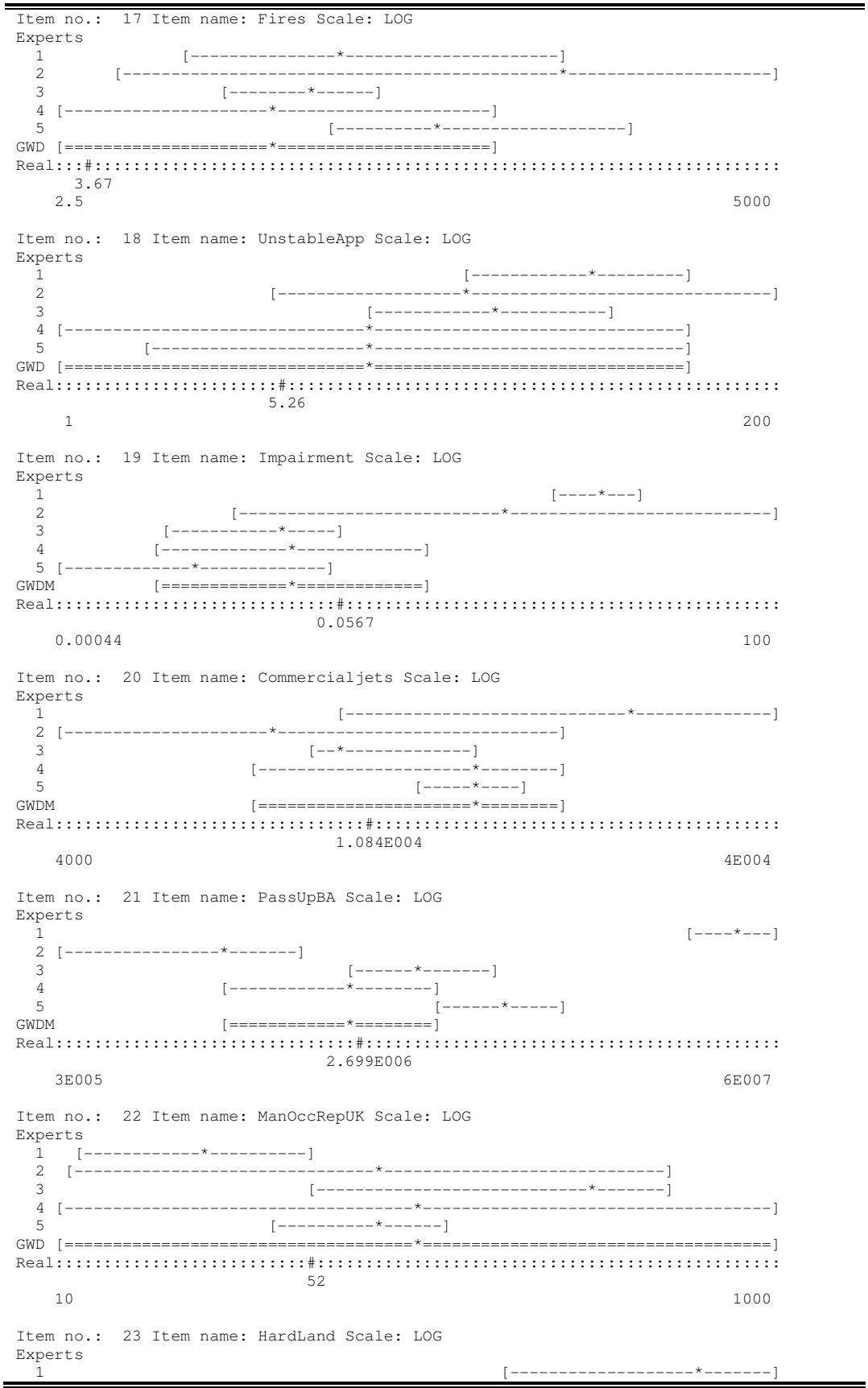


Figure 2. Range Graphs for expert's assessments in variables of interest.

To get a picture of the degree of homogeneity within the expert group, range graphs showing all assessments per item are useful. Figures 2 and 3 show experts' assessments for variables of interest and calibration variables. For each expert, the upper and lower quantiles are given as "[---] ", and the median is given as " * " . The intrinsic range is shown below the expert assessments.


Item no.: 17 Item name: Fires Scale: LOG
Experts
1
2
3
4
5
GWD
Real
3.67
2.5
5000
Item no.: 18 Item name: UnstableApp Scale: LOG
Experts
1
2
3
4
5
GWD
Real
5.26
1
200
Item no.: 19 Item name: Impairment Scale: LOG
Experts
1
2
3
4
5
GWDM
Real
0.0567
0.00044
100
Item no.: 20 Item name: Commercialjets Scale: LOG
Experts
1
2
3
4
5
GWDM
Real
1.084E004
4000
4E004
Item no.: 21 Item name: PassUpBA Scale: LOG
Experts
1
2
3
4
5
GWDM
Real
2.699E006
3E005
6E007
Item no.: 22 Item name: ManOccRepUK Scale: LOG
Experts
1
2
3
4
5
GWD
Real
52
10
1000
Item no.: 23 Item name: HardLand Scale: LOG
Experts
1

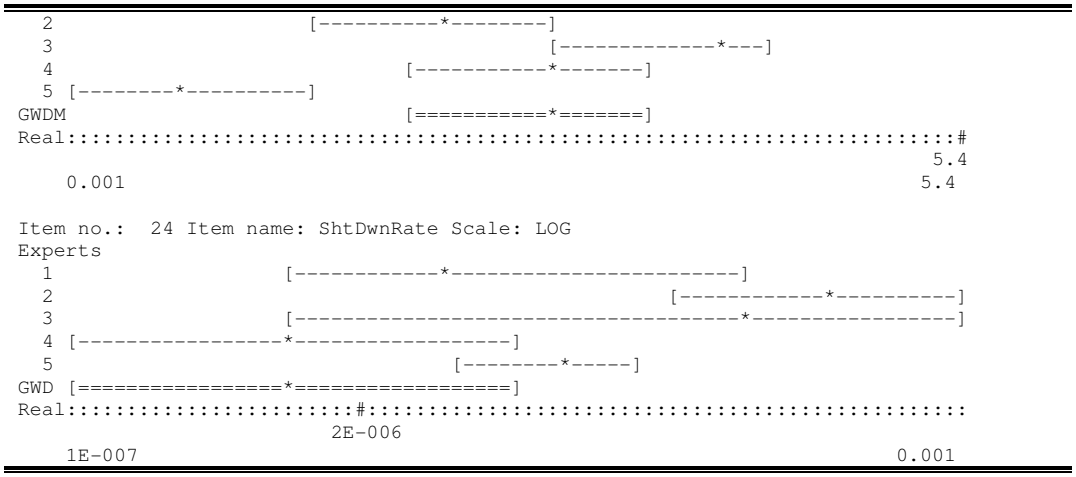


Figure 3. Range Graphs for expert's assessments in calibration variables.

The predominant picture is that the experts' central 90% confidence, generally show considerable overlap. To appreciate Figure 2 numerical table 4 also compares each expert to the GWDM. In discrepancy analysis, the relative information of each expert with respect to the Decision Maker is computed per item. These scores are averaged over all variables (column 10) and calibration variables only (column 11). It can be observed that experts tend to agree with the GWDM less in the calibration questions. The ratio of columns 10 to 11 is as big as 2 and as small as one which supports reasonably the fact that the calibration questions are representative of the variables of interest.

In summary the recommended choice of the decision maker is the Global Weight Decision Maker as it achieves better performance than the Equal weight and Global weight without optimization combinations. The distributions from Figure 1 were obtained from this decision maker and future analysis will be performed also based on the GWDM.

Combining Experts' Dependence Opinions.

As explained before, in the classical method for expert judgment, the marginal distributions of experts e = 1,…,N are combined through a weighted average. In equation (1) $f_{Y,DM}$ is the density of elicitation variable Y of the DM, w_e the weight for expert e and $f_{Y,e}$ the density of expert e on variable Y. The weights can be chosen as equal weights or derived from performance measures.

$$f_{Y,DM} = \sum_{e=1}^{N} w_e f_{Y,e} \tag{1}$$

From previous analysis it was concluded that the GWDM would be the combination choice. This Decision Maker is equal to the opinion of expert *B*. No combination is required for the dependence opinion since the rank and conditional rank correlations will be taken as the opinion of this single expert. The results are presented in table 5. and the fully quantified model in Figure 4.

Rank Correlation	Value	Rank Correlation	Value
$r_{7,1}$	-0.95	$r_{10,7\|6}$	1
$r_{7,3\|1}$	0.86	$r_{13,10}$	0.3
$r_{7,2\|1,3}$	0.24	$r_{13,11\|10}$	-0.32
$r_{6,5}$	-0.95	$r_{13,8\|10,11}$	0.46
$r_{6,3\|5}$	0.86	$r_{13,12\|10,11,8}$	0.18
$r_{6,4\|5,3}$	0.24	$r_{13,9\|10,11,8,12}$	0.19
$r_{10,6}$	0.71		

Table 5. GWDM Dependence Information

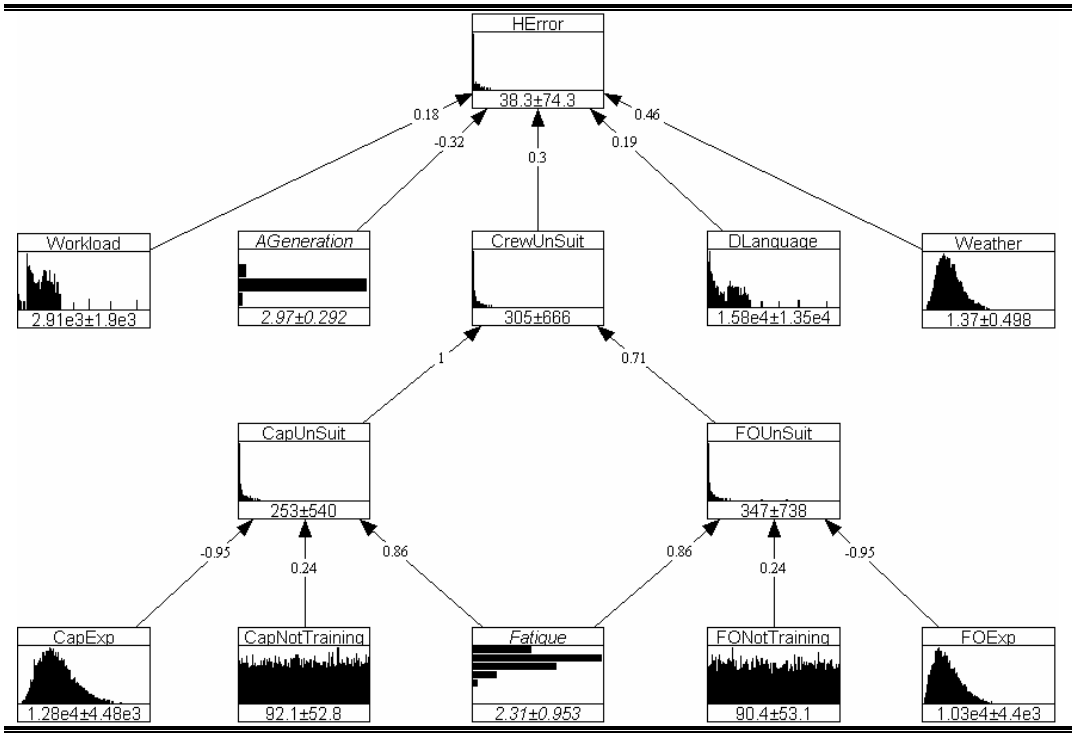


Figure 4. Flight Crew Performance Model.

All rank correlations were computed from conditional probabilities of exceedence as in [2]. Although no combination of the dependence information was performed for this model an explanation about how this combination would be done is presented in appendix 1.

The exceedence probability technique, with the particular requirements of each part in the BBN presented in Figure 4 was employed to obtain all the dependence information presented in Table 2 except for $r_{10,6}$ and $r_{10,7|6}$. The way these two rank correlations were obtained is discussed next.

The rank correlations $r_{10,6}$ and $r_{10,7|6}$ are constrained by the experts' assessments on the variables with index smaller than 6 and 7. These two correlations where chosen such that $r_{10,6}$ and the unconditional correlation $r_{10,7}$ (not the conditional rank

correlation $r_{10,7|6}$ itself) are as big as possible and equal. This is accomplished by choosing $r_{10,6}$ and $r_{10,7|6}$ as in Table5. A more formal presentation of this problem can be found in Appendix 2 at the end of this report.

BASIC BIBLIOGRAPHY.

[1] Cooke R.M. "*Experts in Uncertainty: Opinion and Subjective Probability in Science*" in Environmental Ethics and Science Policy Series, Oxford University Press, June 1991.

[2] Morales Napoles, O. "*Eliciting conditional and unconditional rank correlations from conditional probabilities*" Reliability Engineering & System Safety on press doi:10.1016/j.ress.2007.03.020.

APPENDIX 1.

To combine the dependence information elicited from expert's via conditional probabilities it would be tempting to pool the conditional probabilities linearly to determine the conditional probability of the decision maker. This strategy would work well in general if the medians off all experts were the same which is not typically the case. In order to combine the experts' dependence information a different strategy has to be taken. An example is presented with the BBN formed by variables 7, 3, 2 and 1. A formal description of the technique will be presented in future versions of this report.

(i) Query from expert e_1 the exceedence probabilities (2) and translate them to the corresponding rank correlations

$$
\begin{aligned}
P_1^{e_1} &= P(X_7 \geq x_{7,50} \mid X_1 \geq x_{1,50}) \\
&= P(F_{X_7}^{e_1}(X_7) \geq 0.5 \mid F_{X_1}^{e_1}(X_1) \geq 0.5) \\
P_2^{e_1} &= P(X_7 \geq x_{7,50} \mid X_1 \geq x_{1,50}, X_3 \geq x_{3,50}) \\
&= P(F_{X_7}^{e_1}(X_7) \geq 0.5 \mid F_{X_1}^{e_1}(X_1) \geq 0.5, F_{X_3}^{e_1}(X_3) \geq 0.5) \\
P_3^{e_1} &= P(X_7 \geq x_{7,50} \mid X_1 \geq x_{1,50}, X_3 \geq x_{3,50}, X_2 \geq x_{2,50}) \\
&= P(F_{X_7}^{e_1}(X_7) \geq 0.5 \mid F_{X_1}^{e_1}(X_1) \geq 0.5, F_{X_3}^{e_1}(X_3) \geq 0.5, F_{X_2}^{e_1}(X_2) \geq 0.5)
\end{aligned}
\qquad (2)
$$

$$
\begin{aligned}
P_1^{e_1} &\rightarrow r_{7,1}^{e_1} \\
P_2^{e_1} &\rightarrow r_{7,3|1}^{e_1} \\
P_3^{e_1} &\rightarrow r_{7,2|1,3}^{e_1}
\end{aligned}
$$

Where $F_{X_i}^{e_1}$ is the cumulative distribution function of X_i for expert e_1 and $x_{i,50}$ the median value of X_i. Observe from table one that since X_1, X_2 and X_3 come from data then the median values for these three variables are the same across experts. However, since variable X_7 comes from expert judgment, then the median value across experts and the decision maker will differ.

(ii) Repeat step (i) for experts e_2, e_3, e_4 and e_5.

(iii) Take linear pooling of expert's assessments for X_7 to determine $F_{X_7}^{DM}$ and compute the median value $x_{7,50}^{DM}$ for the Decision Maker.

(iv) Compute the probability that the expert "*would have stated*" if he/she had the same median as the Decision Maker (3) such that his estimates for the rank correlations remain unchanged:

$$
\begin{aligned}
r_{7,1}^{e_1} &\rightarrow P_{1*}^{e_1} \\
r_{7,3|1}^{e_1} &\rightarrow P_{2*}^{e_1} \\
r_{7,2|1,3}^{e_1} &\rightarrow P_{3*}^{e_1} \\
P_{1*}^{e_1} &= P(F_{X_7}^{e_1}(X_7) \geq F_{X_7}^{e_1}(x_{7,50}^{DM}) \mid F_{X_1}^{e_1}(X_1) \geq 0.5) \\
P_{2*}^{e_1} &= P(F_{X_7}^{e_1}(X_7) \geq F_{X_7}^{e_1}(x_{7,50}^{DM}) \mid F_{X_1}^{e_1}(X_1) \geq 0.5, F_{X_3}^{e_1}(X_3) \geq 0.5) \\
P_{3*}^{e_1} &= P(F_{X_7}^{e_1}(X_7) \geq F_{X_7}^{e_1}(x_{7,50}^{DM}) \mid F_{X_1}^{e_1}(X_1) \geq 0.5, F_{X_3}^{e_1}(X_3) \geq 0.5, F_{X_2}^{e_1}(X_2) \geq 0.5)
\end{aligned}
\tag{3}
$$

(v) Repeat step (iv) for expert e_2, e_3, e_4 and e_5.

(vi) Take linear pooling of $P_{j*}^{e_i}$ $i = 1, \ldots, 5$ and j = 1, 2, 3 to determine (4).

$$
\begin{aligned}
P_1^{DM} &\rightarrow r_{7,1}^{DM} \\
P_2^{DM} &\rightarrow r_{7,3|1}^{DM} \\
P_3^{DM} &\rightarrow r_{7,2|1,3}^{DM}
\end{aligned}
\tag{4}
$$

Since in our example all experts except 2 have weight zero then the dependence estimates from this expert will be the only ones to contribute in the linear combination.

APPENDIX 2.

The optimization problem can be stated as follows:

$$\text{A: } \max(r_{10,6}) \text{ s.t.}$$
$$0 \le r_{10,6} \le 1$$
$$0 \le r_{10,7} \le 1$$
$$0 \le r_{10,7|6} \le 1$$
$$r_{10,7} = r_{10,6}$$

An optimization routine was used to cope with the maximization problem 'A' stated previously. Graphically the reader may observe the problem in Figure 2.

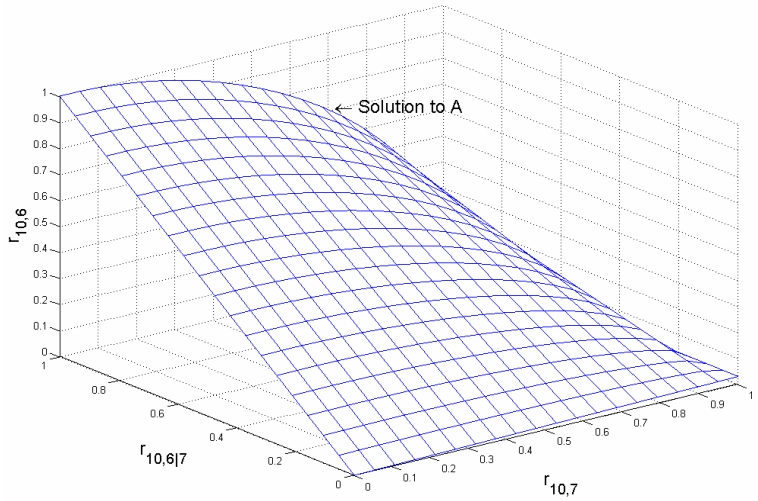


Figure A1. Graphical representation of the maximization problem A.