Expert Elicitation of Low Probability Events: The Roles of Personality and Feedback

Darren Hudson, Combest Endowed Chair,
Texas Tech University
Keith Coble, Giles Distinguished Professor,
Mississippi State University

Presented to the Department of Behavioral Sciences, **United States Military Academy**, March 2014.

Needs for Subjective Risk Assessment

- Most situations of interest have limited data on which to base objective assessments.
- The areas of interest are the tails of distributions, and discernment of low probability events poses problems for many people.
- Obvious military/intelligence applications

Issues with Subjective Assessment

- Methods of elicitation and aggregation
  - Composite
  - Composite with Feedback
  - Consensus
- Concerns about heuristics and biases brought in by experts (Kahneman et al.)
- Personality and risk preferences may influence answers (Chauvin et al.)

Issues with Subjective Assessment

- No real consensus on the proper elicitation methods
  - Real need to rigorously examine elicitation techniques

Objectives

- Examine the impact of elicitation procedure on subjective assessment accuracy and bias
- Examine the role of individual personality type and risk preferences on subjective assessment and accuracy and bias
- Utilize an economic experiment to insure incentive compatibility

Methods

- Generic—recruited students from the undergraduate and graduate populations
- Given a statistical knowledge test
  - General knowledge, 10 questions multiple choice
  - Used for “weighting” expertise
- Myers-Briggs Personality Inventory
- Risk Aversion Experiment

Risk Aversion

Table 1. Risk preference decision sheet.

Question	Option A	Option B	Which option is preferred?
1	10% chance of $10.00, 90% chance of $8.00	10% chance of $19.00, 90% chance of $1.00	
2	20% chance of $10.00, 80% chance of $8.00	20% chance of $19.00, 80% chance of $1.00	
3	30% chance of $10.00, 70% chance of $8.00	30% chance of $19.00, 70% chance of $1.00	
4	40% chance of $10.00, 60% chance of $8.00	40% chance of $19.00, 60% chance of $1.00	
5	50% chance of $10.00, 50% chance of $8.00	50% chance of $19.00, 50% chance of $1.00	
6	60% chance of $10.00, 40% chance of $8.00	60% chance of $19.00, 40% chance of $1.00	
7	70% chance of $10.00, 30% chance of $8.00	70% chance of $19.00, 30% chance of $1.00	
8	80% chance of $10.00, 20% chance of $8.00	80% chance of $19.00, 20% chance of $1.00	
9	90% chance of $10.00, 10% chance of $8.00	90% chance of $19.00, 10% chance of $1.00	
10	100% chance of $10.00, 0% chance of $8.00	100% chance of $19.00, 0% chance of $1.00	

Table 2. Risk aversion coefficient.

No. of safe choices	Range of relative aversion for $U(W)=\frac{W^{1-r}}{1-r}$	Middle point of relative risk aversion	Risk preference classification
0–1	$-1.76^{a}<r^2<-0.93$	−1.365	Highly risk loving
2	$-0.97<r^2<-0.49$	−0.73	Very risk loving
3	$-0.49<r^2<-0.13$	−0.31	Risk loving
4	$-0.13<r^2<0.19$	0.03	Near risk neutral
5	$0.19<r^2<0.48$	0.335	Slightly risk averse
6	$0.48<r^2<0.78$	0.63	Risk averse
7	$0.78<r^2<1.13$	0.955	Very risk averse
8	$1.13<r^2<1.6$	1.365	Highly risk averse
9–10	$1.6<r^2<2.2^{a}$	1.9	Stay in bed

[a]Those two lower and upper bound are subjectively determined.

Using a 10-sided die insures randomness for the respondent. Payoffs are real.

Economic Experiment

- Respondents provided with samples of data drawn from **known** distributions.

Table 1. Four datasets with known population distribution

Range	$[-\infty, 80)$	$[80, 90)$	$[90, 100)$	$[100, 110)$	$[110, 120)$	$[120, +\infty)$
Datasets	Probability					
Dataset 1: sample of 30 drawn from normal population distribution $N(100, 15)$	0.0912	0.1613	0.2475	0.2475	0.1613	0.0912
Dataset 2: sample of 50 drawn from normal population distribution $N(100, 25)$	0.2119	0.1327	0.1554	0.1554	0.1327	0.2119
Dataset 3: sample of 100 drawn from normal population distribution $N(100, 10)$	0.0228	0.1359	0.3413	0.3413	0.1359	0.0228
Dataset 4: sample of 30 drawn from beta population distribution $\beta(6, 2)$ with range of $(0, 133.33)$	0.1624	0.1203	0.1653	0.199	0.2043	0.1487

Economic Experiment

Presented Questions

(1) What is the chance the variable will fall below 80 next period?
(2) What is the chance the variable will fall at or above 80 and below 90 next period?
(3) What is the chance the variable will fall at or above 90 and below 100 next period?
(4) What is the chance the variable will fall at or above 100 and below 110 next period?
(5) What is the chance the variable will fall at or above 110 and below 120 next period?
(6) What is the chance the variable will fall at or above 120 next period?

Payoffs Generated

$$\text{Payoff} = \text{Maximum}\left(\$0.00,\ \$10.00 - 0.025 \times \sum_{i=1}^{6}(A_i - B_i)^2\right)$$

The payoffs are intended to focus the respondent's attention on accuracy, and, thus, maximum payoff. Scalar effects of the size of payoff are always relevant, but at least a weak preference for accuracy is generated by this incentive compatibility measure.

Economic Experiment

- Three treatments
  - Treatment 1—”composite treatment”-respondents provided only individual responses (no feedback), n = 35
  - Treatment 2—” composite with feedback”-respondents received feedback in the form of all individual (anonymous) responses posted on computer screen with opportunity to revise, n = 25 (5 groups of five)
  - Treatment 3—”consensus”-respondents received feedback in the form of all individual (anonymous) responses posted; were paid a $2 bonus of all group members reported the same probability in final round

Analysis

- Two analyses
  - Individual errors (SSE)
  - Aggregated errors (MSE) and biases

Individual Errors

Table 6. Demographic characteristics of participants (continued).

Treatment	Statistical class		Statistical score		Risk preference		Myers–Briggs type indicator							
					No. of safe choice[1]		E		S		T		J	
	Average	SD[a]	Average	SD[a]	Average	SD[a]	Average	SD[a]	Average	SD[a]	Average	SD[a]	Average	SD[a]
1	1.600	1.4142	8.1200	1.2689	5.2500	1.6819	0.60	0.50	0.36	0.49	0.48	0.51	0.60	0.50
2	1.600	2.1602	8.4400	1.4457	4.6842	1.4550	0.44	0.51	0.36	0.49	0.52	0.51	0.52	0.51
3	1.200	0.9129	7.1200	1.4236	5.0588	1.9834	0.60	0.50	0.48	0.51	0.60	0.50	0.52	0.51
Total	1.467	1.5711	7.8930	1.4757	5.0000	1.6949	0.55	0.50	0.40	0.49	0.53	0.50	0.55	0.50

[a]SD=standard deviation.

[1]We drop those subjects who switched from B back to A. So the valid observations are 20, 19 and 17 in Treatment 1, 2 and 3, respectively.

Individual Errors

$$\mathrm{EIE} = \beta_1 + \beta_2 \mathrm{TD}_2 + \beta_3 \mathrm{TD}_3 + \beta_4 \mathrm{DD}_2 + \beta_5 \mathrm{DD}_3 + \beta_6 \mathrm{DD}_4 + \beta_7 \mathrm{MD}_1 + \beta_8 \mathrm{MD}_2 + \beta_9 \mathrm{MD}_3 + \beta_{10} \mathrm{MD}_4 + \mathrm{STAT} + \mathrm{RISK} + \varepsilon$$

- TDs are the treatment effects (composite base)
- DDs are the dataset effects (small sample normal base)
- MDs are MBPI designations
  - MD1 = 1 if extrovert
  - MD2 = 1 if sensing
  - MD3 = 1 if thinking
  - MD4 = 1 if judging

Individual Errors

Table 7. Regression output for the error of individual estimate model.

Variables	Error of individual estimate across entire distribution		Error of individual estimate across left tail		Error of individual estimate across right tail		Error of individual estimate across two tails	
	$(-\infty, +\infty)$		$(-\infty, 80)$		$(120, +\infty)$		$(-\infty, 80)$ and $(120, +\infty)$	
	Coefficient	p value	Coefficient	p value	Coefficient	p value	Coefficient	p value
Intercept	861.6293	<0.001***	85.9054	**<0.001*** **	141.2804	**0.0072*** **	207.1858	**0.0005*** **
TD_2	−97.0366	0.0091***	−8.2458	0.0826*	−25.3732	0.1171	−33.619	**0.0673***
TD_3	−114.5273	0.0105	−10.5297	**0.0657**	−26.1358	0.1796	−36.6655	**0.0971***
DD_2	−437.1964	**<0.001*** **	−0.1559	**0.9772**	−0.1195	0.9949	−0.2755	0.9896
DD_3	−435.9869	**<0.001*** **	−14.6464	0.0077***	−2.8851	0.8779	−17.5015	0.4066
DD_4	−195.6332	**<0.001*** **	−1.5941	0.7804	24.1884	0.2156	22.5944	0.3073
MD_1	−15.3605	0.6323	−1.2013	0.7704	−27.4404	**0.0516***	−28.6477	**0.0729***
MD_2	8.5139	0.8005	8.7103	**0.0448** **	39.2367	**0.0083*** **	47.947	**0.0045*** **
MD_3	42.5189	0.223	−3.1618	0.4793	−2905273	**0.0536***	−32.6891	**0.0594***
MD_4	−87.014	**0.0078*** **	−5.8462	0.1608	−24.2048	**0.089***	−30.051	**0.0627***
STAT	−25.7053	**0.0498**	−4.5868	**0.0066*** **	−8.9195	0.1189	−13.5063	**0.0377**
RISK	7.7996	0.7322	−3.0151	0.412	−2.2178	0.8594	−5.2329	0.7125
No. of Obs	224		224		224		224	
R^2	0.4502		0.1249		0.106		0.1266	

***, ** and * denote significance at the 1, 5 and 10% levels, respectively.

Individual Errors

- Feedback improves accuracy
  - Means of feedback seems less important than the existence of feedback
- Distribution/sample size appears not to significantly affect error rates
- Knowledge/experience matters
  - Limited in scope here, but supports the idea that true experts produce better subjective estimates

Individual Errors

- Personality effects
  - Extroverts performed better—focus attention to outer world of people and things; perhaps outward orientation leads to better induction?
  - Sensing were more inaccurate—focus attention on five senses and discount "gut feelings"; intuitive people listen to their unconscious mind
  - Thinkers performed better—focus attention on facts and objective reasoning rather than person-centered emotional responses
  - Judgers performed better—prefer planned and organized approach to life, likely leads to more organized thought processes

Personality matters…now you just have to figure out how to effectively screen for it with your experts.

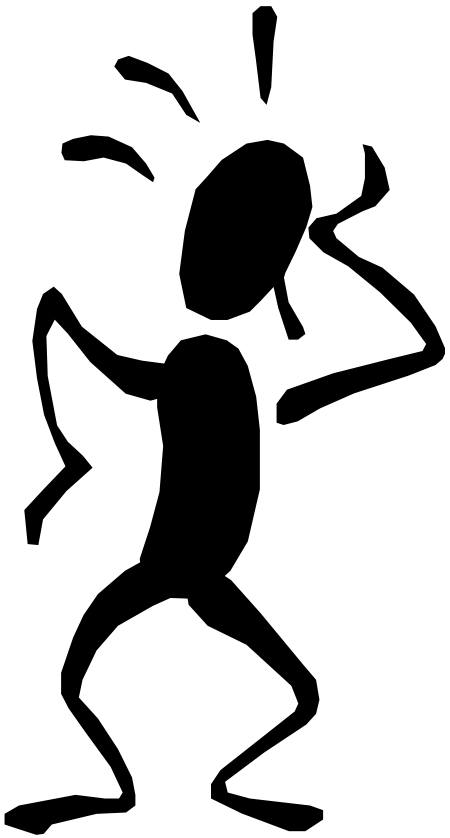
Aggregate Errors

Table 3. Regression output for the MSE of aggregated estimate model

	MSE of aggregated estimate across left tail		MSE of aggregated estimate across right tail		MSE of aggregated estimate across two tails	
	$(-\infty, 80)$		$(120, +\infty)$		$(-\infty, 80)$ and $(120, +\infty)$	
Variables	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value
Intercept	16.7026	<0.0001***	23.661	<0.0001***	71.389	<0.0001***
TD2	−1.2681	0.4256	−7.0779	0.0005***	−9.5258	0.0469**
TD3	−2.9321	0.0763*	−5.1874	0.0061***	−12.6912	0.011**
DD2	−8.621	0.0002***	−3.4897	0.0858*	−24.1839	0.0002***
DD3	−13.9041	<0.0001***	−15.2992	<0.0001***	−57.1299	<0.0001***
DD4	−2.5515	0.173	6.366	0.004***	5.3062	0.3157
WD	−0.1949	0.8797	−0.1199	0.9304	−0.1132	0.9755
No. of obs.	24		24		24	
R^2	0.8178		0.9013		0.9192	

Note: ***, ** and * denote significance at the 1, 5 and 10% levels, respectively.

Aggregate Errors



- Again, feedback matters, but it does not seem to matter the form of the feedback.
- Here, sample size does matter and reduces aggregate error rates
- Weighting scheme does not seem to matter

Aggregate Bias

Table 4. Regression output for the bias of aggregated estimate model

	Bias of aggregated estimate across left tail		Bias of aggregated estimate across right tail		Bias of aggregated estimate across two tails	
	$(-\infty, 80)$		$(120, +\infty)$		$(-\infty, 80)$ and $(120, +\infty)$	
Variables	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value
Intercept	3.2218	<0.0001***	4.6062	<0.0001***	7.828	<0.0001***
TD2	0.7805	0.005***	−0.4013	0.2114	0.3792	0.4555
TD3	0.2199	0.3771	−0.6497	0.0507*	−0.4297	0.3987
DD2	−5.5765	<0.0001***	−7.9566	<0.0001***	−13.533	<0.0001***
DD3	−4.3268	<0.0001***	−3.5692	<0.0001***	−7.8961	<0.0001***
DD4	−0.6108	0.0435*	0.4798	0.1964	−0.131	0.8219
WD	0.0339	0.866	−0.0259	0.9195	0.008	0.9844
# of Obs	24		24		24	
R^2	0.9718		0.977		0.9789	

Note: *** and * denote significance at the 1 and 10% levels, respectively.

Aggregate Bias

- No real consistent pattern in bias relative to feedback
  - Perhaps the consensus reduced it a bit, but the relationship is not particularly strong
- Bigger samples help reduce bias, but not clear if there is an “optimal” sample size
- Weighting scheme does not matter

Conclusions

- Use of incentive compatible experiment reveals some useful information:
  - Feedback is useful, but the manner is less important
    - If aggregating with simple feedback is cheaper than a “Delphi” approach, it may be preferred
  - Sample size matters in the aggregate, but no so much at the individual level
  - Personality and experience or knowledge matter, but risk aversion does not
  - No consistent pattern to bias
    - We did not allow non-anonymous interaction where personality and bias may interact; could be a real weakness of “Delphi” approaches

Recommendations

- Any attempt to elicit subjective estimates of risk should include:
  - Some manner of interactive/iterative feedback
  - Persons screened for specific expertise
  - Persons screened for personality types
  - Attention to sample size…but, more attention to feedback