Using Expert Judgment to Understand the Rare Event Threat Space of Homeland Security: Practices, Challenges, and Opportunities

Paul S. Szwed
U.S. Coast Guard Academy
Paul.S.Szwed@USCGA.edu
11.14.2011

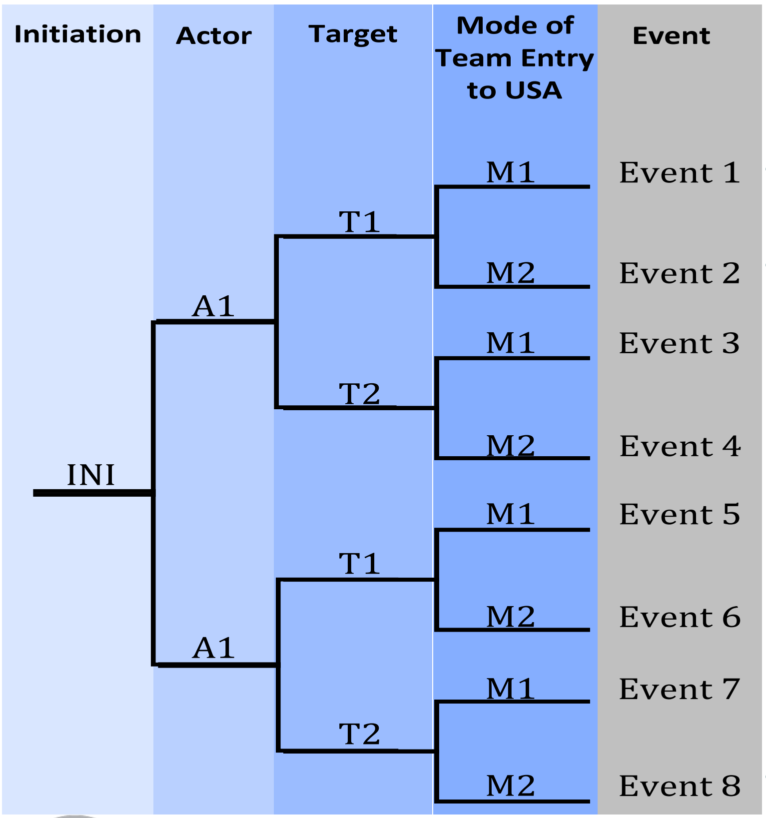
Homeland Security

Irving Susel, Natasha Hawkins, Debra Elkins, Antonio Kirson
Office of Risk Management & Analysis
Department of Homeland Security

Background

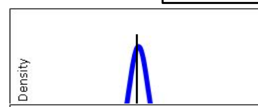
- Homeland Security Presidential Directives 10, 22, and 18
- Quadrennial Homeland Security Review & Bottom-up Review
- DHS Policy for Integrated Risk Management (June 27, 1010)

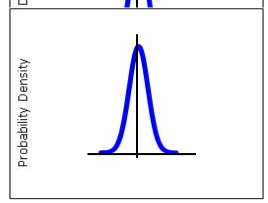
Terrorism Risk Assessment



Iteration	Sampled Probability
1	0.25

Iteration	Sampled Probability
1	0.75
2	0.72
...	...
500	0.67

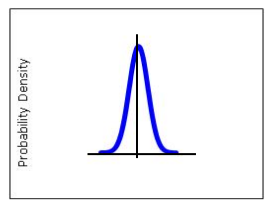




⋮

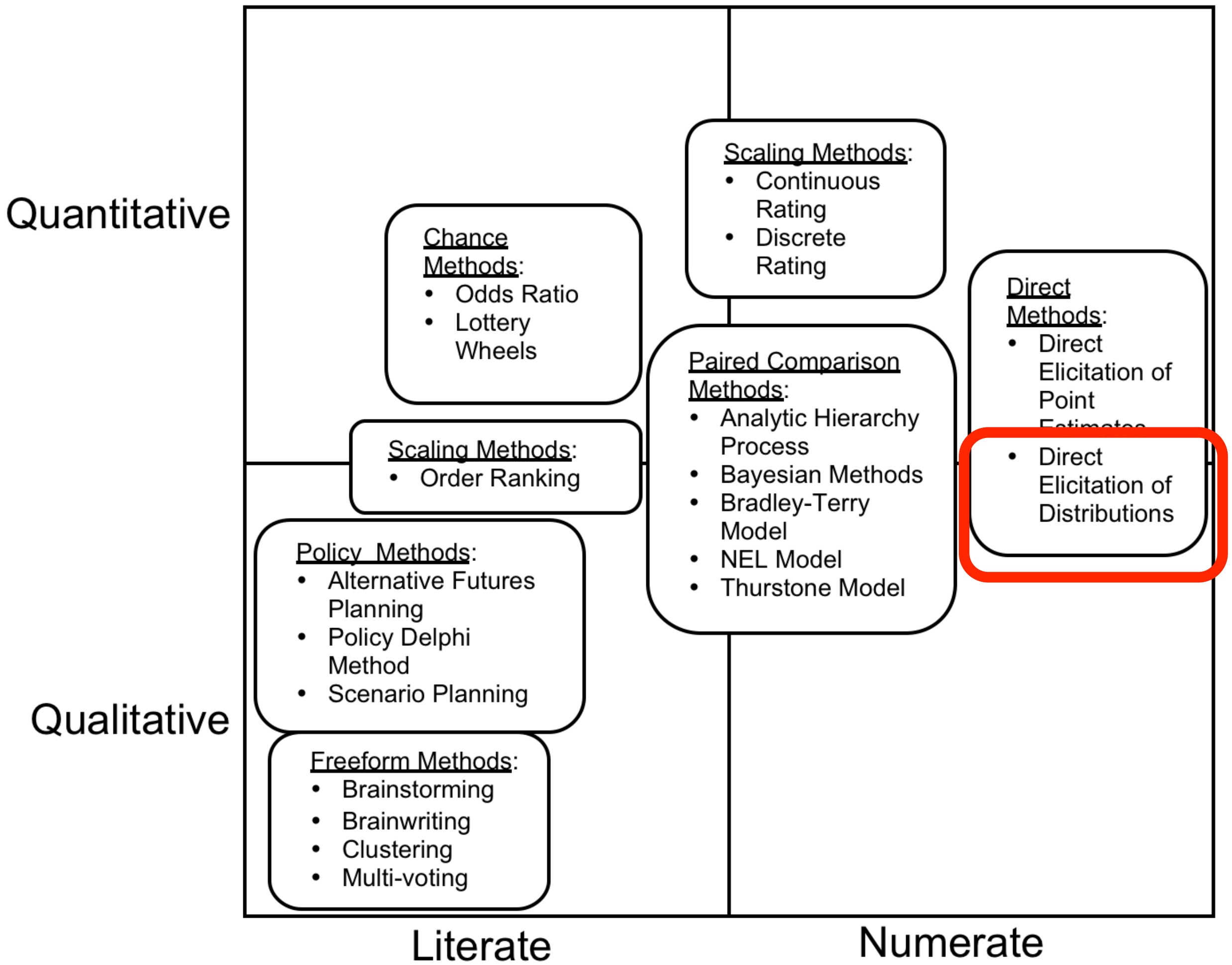
⋮

Iteration	Sampled Probability
1	0.12
2	0.13
...	...
500	0.14



Homeland Security

Expert Judgment



NUREG-1150


Eliciting Probabilities from Experts in Complex Technical Problems

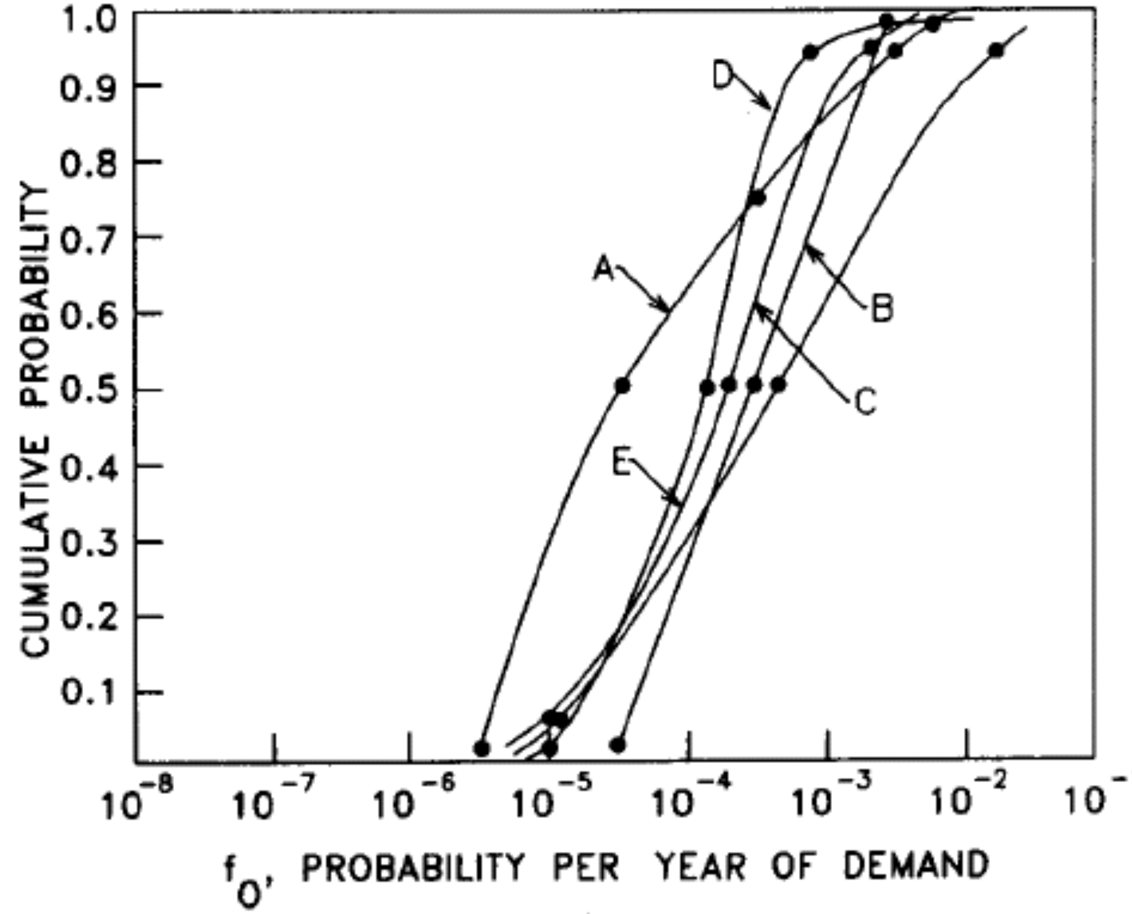


Fig. 4. Cumulative probability distributions for five experts over probability of failure of valve to open.

Homeland Security

DHS Expert Elicitation Protocol

- **Identify issues and select experts**
- **First Meeting**
  - Discuss issues, share knowledge
  - Define variables and events – “elicitation statement”
  - Probability Training
  - Practice
- **Individual Elicitations**
- **Additional questions completed by survey**
- **Aggregation and Documentation**
  - Aggregate expert judgments
  - Document substantive reasoning
  - Document probability reasoning
- **“Wrap-up” Meeting**
  - Review findings, share knowledge
  - Review and reconciliation

Elicitation Tool Screenshot (Example)

What is the expected/estimated relative frequency with which **International Terrorist Groups** would select each of the four agents being considered?

Type of Attack Scenario	Short Name	Rank, High to Low (1=Highest)	Relative to most likely (most likely = 1)	Inverse Ratios	Probability
Chemical Agent B	Agent B	1	1	1	**81.07%**
Chemical Agent C	Agent C	2	5	0.2	**16.21%**
Chemical Agent A	Agent A	3	30	0.033333333	**2.70%**
Chemical Agent D	Agent D	4	10000	0.0001	**0.01%**
			Sum of Inverse Ratios	1.23	

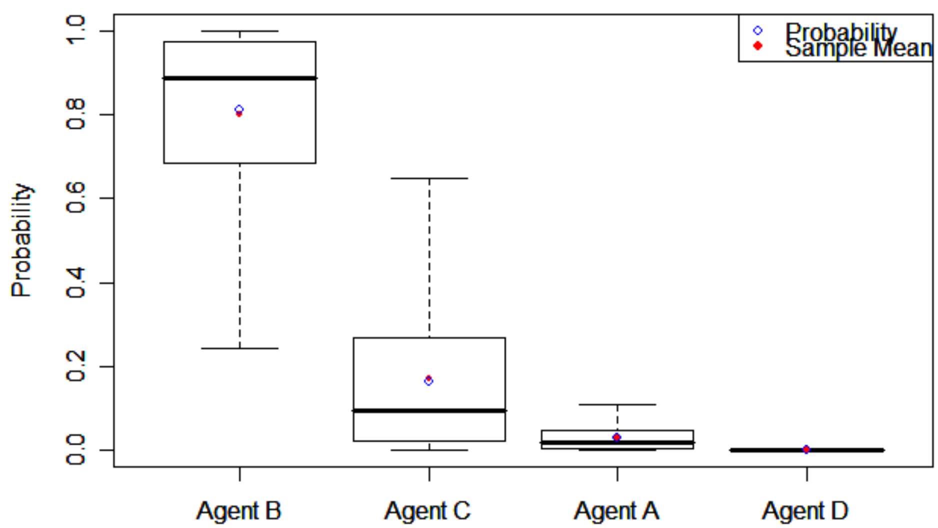
K Estimation	
Most Likely Probabilty	0.8107
5th Percentile	0.2
95th Percentile	0.9
	K
Range Calculation	3
Upper Calculation	39
Lower Calculation	1

Key Details to Keep in Mind

For this discussion, **International Terrorist Groups** are defined as any terrorist group that operates both inside and outside the U.S. and is not sponsored by a nation.

Only consider attacks on the U.S. Homeland (not U.S. interests abroad).

Timeframe of interest is 2008-2012.



Data randomly generated – not actual elicitation data

Challenges & Opportunities (1)

Plan Elicitation

- Requisite Expertise
- Size of Expert Pool
- Composition of Pool
- Frequency of Elicitation
- Form of Judgments
- Protocol Testing
- Elicitation Mode
- Elicitation Interface
- Amount of Interaction
- Expert Fatigue

Select & Train Experts

- Availability of Experts
- Criteria for Selection
- Probability Training
- De-biasing Training
- Protocol (instrument) Familiarization
- “Composite” Judgments

Homeland Security

Challenges & Opportunities (2)

Elicit Judgments

- Mode of Elicitation (Direct vs. Relative)
- Expert Interface
- Dealing with Uncertainty
- Use of Prior Judgments
- Order of Elicitation
- Feedback to/from Experts
- Individual vs. Group
- Description of Judgments (e.g. shape of distribution)

Analyze & Aggregate Judgments

- Calibration of Experts (e.g. performance)
- Weighting of Experts (e.g. scoring rules)
- Dealing with Lack of Consensus
- Aggregation?
- Method of Aggregation
- Sensitivity Analysis
- Multiple-methods

Homeland Security

Challenges & Opportunities (3)

Document & Communicate Results

- Transparency vs. Sensitivity
- Capture Expert Rationale
- Peer Review of Judgments / Model
- Validation of Model
- Decision Making Process / Stakeholder Requirements
- Data Display / Visualization
- Risk Preference or Tolerance
- Communicating Uncertainty

Elicitation of Expert Judgment – Program Effectiveness Judgments

Study 1: Interval Estimation

- Which mode of expert elicitation is most effective in reducing overconfidence bias?

Study 2: Decomposition

- What sort of decomposition is most effective in improving accuracy and confidence of expert judgments?

RAPID 2010: modified Kent Scale due to time constraints

- Effectiveness judgments were required for each activity, or node, within an incident chain where programs directly manage risk, and are based upon the unique roles they perform at different points on an incident chain:
  - Detect: What is the overall likelihood that the program's resources/assets would detect or find the adversary, weapons materials, or illegal goods?
  - Interdict: What is the overall likelihood that the program's resources/assets would successfully interdict the adversary, weapons materials, or illegal goods?
  - Protect: What percentage of the potential damage caused by an attack/hazard would the program's resources/assets preemptively mitigate?
  - Respond: what percentage of acute localized injuries that would result in death if untreated would the program successful treat after an attack/hazard has occurred?
  - Recover: what percentage of localized, chronic medical conditions that would result in death in the long term if untreated would the program successfully treat AND/ORwhat percentage of immediate national and subsequent national and local economic losses would the program mitigate after an attack/hazard has occurred?

RAPID II Effectiveness Scale – Operational Programs

Likelihood	Synonym	Percent	Range
Certain	Absolute; Authoritative; Clear; Conclusive; Confirmable; Definite	100%	100
Nearly certain	Virtually (almost) certain; We are convinced; Highly probable; Highly likely	93%	87-99
Probable	Likely; We believe; We estimate; Chances are good; It is probable that	75%	60-86
Even	Chances are even; 50-50	50%	37-59
Improbable	Probably not; Unlikely; We believe… not	30%	14-36
Nearly impossible	Almost impossible; Only a slight chance; Highly doubtful	7%	2-13
Practically impossible	Absurd; Infeasible; No-way; Preposterous	1%	1

Homeland Security

Overconfidence in Interval Estimation:
Survey of State of the Science (past decade, or so)

Well known that experts are subject to judgmental biases when assessing subjective probabilities (e.g. Khaneman et al. 1982).

Prevalence of overconfidence in assessing probabilities noted in *Calibration of Probabilities: The State of the Art to 1980* (Lichtenstein et al., 1981).

2004

- 3-point estimates > 2-point estimates > Range estimates (Soll & Klayman)
- Interval production > Interval evaluation (Winman, Hansson, & Juslin)
- Lack of feedback, perspective, hidden agenda, and reliance on estimator protocols may be factors in overconfidence (Jørgensen, Teigen, & Moløkken)

2005

- Participant assigned interval > Analyst assigned interval + Feedback improves calibration (Teigen & Jørgensen)
- Miscalibration reduces & self-monitoring increases performance (Biais, Hilton, Mazurier & Pouget)

2008

- While experts provide narrower intervals and midpoints nearer *true* values, experts and lay people exhibit similar calibration (McKenzie, Liersch, & Yaniv)
- Asking most likely first causes anchoring in lay people, but less so in experts (Bruza, Welsch, Navarro, & Begg)
- Significant differences in overconfidence were found across 27 different expert panels (~5,000 estimations including many repeated under different conditions), but the differences were not attributable to question effect (Lin & Bier)

Wisdom of Crowd in One Mind:

2008: Average of two estimates more accurate than either estimate + time delay between estimates improves average (Vul & Pashler)

2009: Average improves using dialectic (consider the opposite) estimate (Herzog & Hertwig)

2010: Averaging own estimates improves own accuracy only when accessing different knowledge, but does not outperform average with another participant (Rauhut & Lorenz)

2009

- More-Or-Less Estimation (MOLE) outperforms interval estimation techniques (Welsh, Lee, & Begg)
- When provided others estimates as feedback, participants tended to choose their or the others, but it would be more effective to average (Soll & Larrick)
- Upper limits preferred for small values, lower limits for larger values or as default (Halber & Teigen)

2010

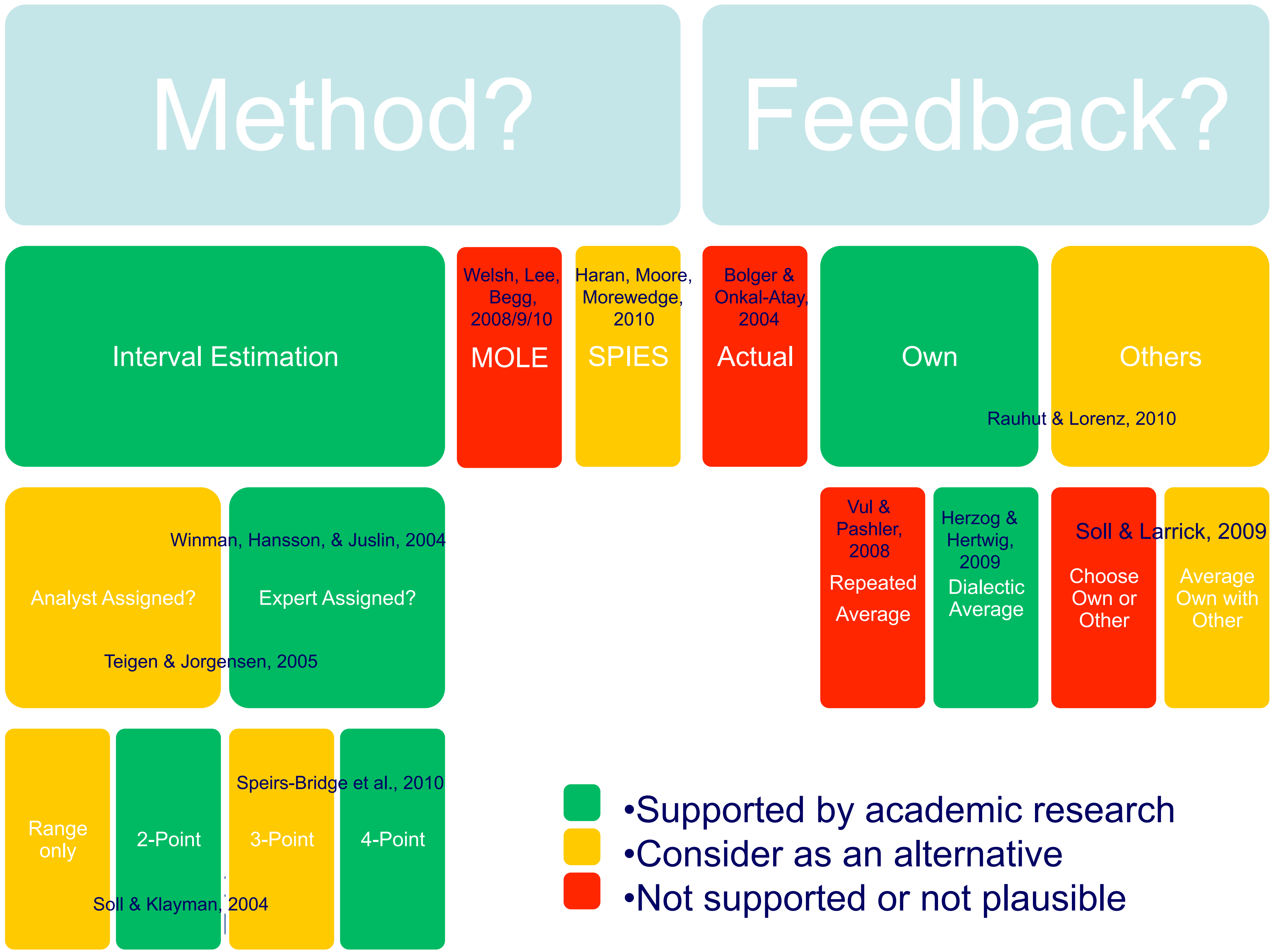
- Requiring participants to provide estimates of each bin within a range of all possible outcomes increases precision (Haran, Moore, & Morewedge)
- Calibration increases immediately after low-probability, high-consequence and increases with time (Li et al.)
- 4-point estimate > 3-point estimate (Speirs-Bridge et al.)

2011

- Providing more proximal intermediate estimates ("unpacking the future") improves calibration (Bearden, Gaba, & Mukherjee)

NOTE: A meta-search of over 100 bibliographic databases identified 2,092 articles with "overconfidence" in the title. Of these, 16 (some with duplicates) clustered on "interval estimates." The 20 shown here include those identified in the search.

Homeland Security


Method?
Feedback?
Interval Estimation
Welsh, Lee, Begg, 2008/9/10
MOLE
Haran, Moore, Morewedge, 2010
SPIES
Bolger & Onkal-Atay, 2004
Actual
Own
Others
Rauhut & Lorenz, 2010
Winman, Hansson, & Juslin, 2004
Analyst Assigned?
Expert Assigned?
Teigen & Jorgensen, 2005
Vul & Pashler, 2008
Repeated Average
Herzog & Hertwig, 2009
Dialectic Average
Soll & Larrick, 2009
Choose Own or Other
Average Own with Other
Speirs-Bridge et al., 2010
Range only
2-Point
3-Point
4-Point
Soll & Klayman, 2004
•Supported by academic research
•Consider as an alternative
•Not supported or not plausible

No Decomposition Baseline

What's the likelihood of detection, keeping in mind:

- whether or not vessels are likely to be boarded,
- if you board, the likelihood that you will detect terrorists that are there
- all of the factors that contribute to this.

This will be estimated to two ways:

Overall Likelihood of Detection

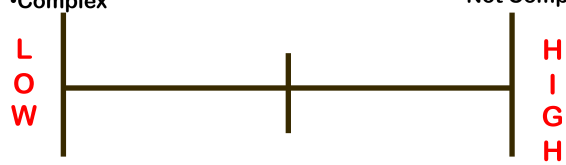
Selecting the first, second and third choice Kent Scale bins.

Overall Likelihood of Detection

Estimating a 90% confidence interval, describing the circumstances for the low, high and likely estimates.

- No intel inject
- Not Boarded
  or
- Large Vessel
- Complex

- Specific Intel Inject
- Boarded
- Small Vessel
- Not Complex



The range reflects both:

- Remaining Variability
and
- Uncertainty

Homeland Security

3-point Estimate + Decomposition

Factor

Intel Inject
- No Intel
- General
- Specific

Size of Vessel
- Small (<300GT)
- Large (>300 GT)

Harbor Type
- Low Traffic
- High Traffic

High Interest Vessel?
- Non-HIV
- HIV

Vessel Complexity
- Low Bulk Cargo
- High Container Ship

Homeland Security

NEW APPROACH FOR A PEJ

NOTIONAL DATA

UNCERTAINTY

Effectiveness

Adversary Preference	Incidence	VARIABILITY	Low	Mode	High
10%	20%	**Best Case** *Day & good weather…*	10%	40%	60%
70%	50%	**"Typical" Case** *Detailed description*	5%	10%	25%
20%	30%	**Worst Case** *Night & bad weather…*	0%	5%	10%

Additional INTEL Input?

Homeland Security