Aldo Benini
Patrice Chataigner
Nadia Noumri
Nic Parham
Jude Sweeney
Leonie Tax

A study for ACAPS

The Use of Expert Judgment in Humanitarian Analysis –

*Theory, Methods, Applications*

August 2017
Suggested citation:


Contact information:

ACAPS
23, Avenue de France
CH-1202 Geneva
Switzerland
info@acaps.org
Preface

Experts are indispensable in modern organizations. The humanitarian realm is no exception. Humanitarian experts reduce uncertainty at both extremes – when there are not enough good data, their informed estimates bridge gaps – when there are too many data, they select the variables and models that are the most pertinent. Experts help analysts and responders make sense of scarcity as well as of profusion, by pursuing “What’s the story here?” and “Now what shall we do?” questions.

ACAPS’ reason for being is to strengthen analytic competence in the humanitarian community, including the ability to generate good expert judgment. Expert judgment is a recognized, mature research methodology. But its reach and depth in humanitarian decision-making have not been fathomed out coherently. This study maps the process of expert judgment in this particular task environment. It speaks primarily to humanitarian analysts who oversee the production of expert judgment, but is instructive also for decision-makers, for the experts themselves as well as for interested stakeholders.

The approach is two-pronged. Descriptively, ten case studies exemplify ACAPS’ reliance on expert judgment, proactive and guarded at the same time. The prescriptive section singles out an elicitation process and a number of typical challenges, particularly in the collection and aggregation of this kind of information. It demonstrates solutions that the reader may study and adopt (and improve on) as and when needed. The empirical and technical chapters are bracketed by brief historic and futuristic reflections.

It has been said that expert judgment “is cheap, plentiful, and virtually inexhaustible”. That is, at best, an exaggeration. Experts can be better, faster, or cheaper than other sources and methods. For that to happen, committed decision-makers, competent analysts and sympathetic stakeholders are needed. ACAPS hopes that this study will strengthen their hand and through them enhance the use of humanitarian expert judgment.

Lars Peter Nissen
Director, ACAPS

Geneva, 10 July 2017
**Summary table of contents**

1. Summary ............................................................................................................1
2. Introduction ...................................................................................................... 13
3. The process of expert judgment ................................................................. 32
4. Aggregation and synthesis ........................................................................ 57
5. Humanitarian applications .......................................................................... 115
6. Conclusions and outlook .......................................................................... 169

Appendices ............................................................................................................. 173

References .............................................................................................................. 188
## Detailed contents

1. **Summary** ............................................................................................................1

   1.1. About this study .......................................................................................... 1  
   1.2. Background, preparation, recruitment ........................................................ 2  
   1.3. Elicitation and recording ............................................................................. 4  
   1.4. Aggregation and synthesis ........................................................................... 5  
   1.5. Humanitarian applications ........................................................................... 7  
   1.6. Conclusion .................................................................................................. 9  
   1.7. The way forward ....................................................................................... 10  

2. **Introduction** ...................................................................................................... 13  

   2.1. What this is about ...................................................................................... 13  
   2.2. From Machiavelli to nuclear power ........................................................... 13  
   2.3. Modern expertise and its discontents ......................................................... 15  
   2.4. Expert, analyst, decision-maker ................................................................. 16  
   2.5. Humanitarian experts ................................................................................ 17  
       Are they experts? .............................................................................................. 17  
       A motivating example....................................................................................... 18  
       Yes, they are experts ........................................................................................ 20  
   2.6. Differences and commonalities .................................................................. 23  
       Between humanitarian experts and non-experts................................................. 23  
       Between humanitarian and other experts ........................................................ 26  
   2.7. Summing up .............................................................................................. 30  

3. **The process of expert judgment** ........................................................................ 32  

   3.1. Background and preparation ...................................................................... 34  
   3.2. Recruiting experts ..................................................................................... 37  
   3.3. Eliciting expert judgment .......................................................................... 39  
       A special type of communication...................................................................... 39  
       [Sidebar:] Rapid identification of mine-contaminated communities .......... 40  
       Elicitation Methods........................................................................................... 43  
       [Sidebar:] Digital argument Delphi technique ................................................... 44  
       [Sidebar:] Eliciting judgments about a proportion............................................. 49  
       Conducting the elicitation ................................................................................. 53  
       How much works for humanitarian experts? ..................................................... 55  

4. **Aggregation and synthesis** ................................................................................ 57  

   4.1. Overview .................................................................................................. 57  
   4.2. Synthesis of qualitative expertise ............................................................... 58  
       Step 1: Collection, abstraction, ordering of findings; frequencies ................. 59  
       Step 2: Reorganizing the findings, comparing them in multiple ways ........... 60
Step 3: Extracting the contributors’ own syntheses, importing external concepts............................................................................................................ 62
[Sidebar:] Addressing paradoxes with higher-level interpretation................62
What three-step synthesis achieves … and what not.................................63
[Sidebar:] Research synthesis on the impacts of cash transfers…………………64

4.3. Aggregation of quantitative expert judgments........................................65
Multiple-expert estimates of a scalar.................................................................66
[Sidebar:] Technicalities of aggregating triangular distributions......................71
[Sidebar:] Technicalities of the simplified Beroggi-Wallace method ………………77
Multiple-expert estimates of a proportion........................................................81
[Sidebar:] Technicalities of Bordley’s formula.................................................82
[Sidebar:] The proportion of IDPs in Aleppo, Syria, summer 2015………………..86

4.4. Expert judgment and Bayesian reasoning ..................................................89
Bayes’ theorem .................................................................................................90
[Sidebar:] Numeric demonstration of Bayes theorem ........................................92
Updating beliefs on the strength of new evidence............................................95
A probability scale with unequal intervals.......................................................99
Process tracing and cause-effect testing .........................................................103
[Sidebar:] Belief networks: Migration patterns in the Sahel................................109

5. Humanitarian applications ........................................................................115

5.1. Introduction - The Analysis Spectrum ......................................................115
5.2. Exploratory analysis: Using experts to find information .......................116
When using experts is better, faster and cheaper..........................................116
Application and challenges ..........................................................................117
Lessons learned ..............................................................................................119
5.3. Descriptive analysis: The population inside Syria, 2015 ..........................119
Case: The population inside Syria in summer 2015 .......................................119
Lessons learned ..............................................................................................123
5.4. Explanatory analysis: Problem tree Ebola ...............................................125
Widening the perspective................................................................................125
The ACAPS Ebola project .............................................................................125
The graphic device: The problem tree ..........................................................126
The problem tree as advocacy tool ................................................................130
Lessons learned ..............................................................................................132
5.5. Interpretative analysis: Severity ratings in complex emergencies ..........133
A severity scale for northern Syria.................................................................133
Combining expertise for assessment quality .................................................133
Lessons learned ..............................................................................................134
5.6. Interpretative analysis: Clarifying priorities ..........................................136
Coordination in the early stages of the disaster response .............................136
Collating and reviewing information ............................................................136
Gathering expert perspectives ......................................................................138
Lessons learned ..............................................................................................139
5.7. Interpretative analysis: Information-poor environments .....................140
Mapping information gaps in post-hurricane Haiti .....................................140
ACAPS information gaps methodology in the context of Haiti ..................140
Results............................................................................................................143
Tables

Table 1: Elicitation components and subcomponents .......................................................... 44
Table 2: Different response modes .................................................................................. 48
Table 3: Proportions estimated by four experts ............................................................... 50
Table 4: Combined estimate from four expert judgments .................................................. 51
Table 5: Conducting the interview and moderating the group - step by step .................... 54
Table 6: Example of a taxonomy for re-ordering findings – Segment .............................. 61
Table 7: Arranging findings around particular contributors' conceptual syntheses - Segment ................................................................. 62
Table 8: A numeric example of the simplified Beroggi-Wallace method .......................... 80
Table 9: Indicators of expert quality, Beroggi-Wallace method ........................................ 81
Table 10: Spreadsheet implementation of the simplified Bordley's formula ................. 84
Table 11: An example with one expert providing a very high estimate ............................ 84
Table 12: Population and IDP estimates for a sub-district in Syria, 2015 ....................... 87
Table 13: Aggregation of estimated IDP proportions ....................................................... 88
Table 14: Hypothetical population figures that satisfy Bayesian example ..................... 91
Table 15: Hypothetical population figures with algebraic variable names ..................... 93
Table 16: Medow and Lucey's probability categories ...................................................... 99
Table 17: Process tracing test for causal inference .......................................................... 104
Table 18: Expert judgment and expert consensus on the probability of the evidence ........ 107
Table 19: Estimate of the national population in 2015 ................................................... 121
Table 20: Confidence intervals around the population estimate .................................... 122
Table 21: Segment of a problem tree on secondary impacts of a large-scale Ebola outbreak ................................................................................................................. 130
Table 22: The severity scale used in the J-RANS II ....................................................... 133
Table 23: A seven-level severity rating scale ................................................................. 135
Table 24: Haiti information gap scoring system ............................................................. 141
Table 25: Haiti information gap database (segment) ..................................................... 143
Table 26: Haiti - Cross-tabulation of the analytical value of information, two sectors ...... 144
Table 27: Risk Analysis impact scale ............................................................................. 148
Table 28: Problems faced in the ACAPS Risk Analysis ................................................. 149
Table 29: Steps of the scenario-building process ............................................................ 152
Table 30: Aggregating experts' belief intervals - Simulated example ............................. 158
Table 31: Equal vs. unequal probability ranges .............................................................. 162
Table 32: Examples of calculated functions of the triangular distribution ..................... 179
Figures

Figure 1: The process of expert judgment ................................................................. 2
Figure 2: Updating the probability of new outbreaks ......................................... 19
Figure 3: Ideal-type phases of the expert judgment process .......................... 33
Figure 4: Estimates of mine-affected communities in Thailand, 2000-01, by province .......................................................... 41
Figure 5: The three simultaneous steps of the Argument Delphi method .... 45
Figure 6: Three levels of outcomes in cash transfers ........................................... 65
Figure 7: Two ways for the decision-maker to aggregate the uncertainty of experts 67
Figure 8: Minimum, best estimate, and maximum represented by a triangular distribution .............................................................................. 69
Figure 9: Aggregation of key informant estimates .............................................. 70
Figure 10: Random draws, arranged in a sieve .................................................. 73
Figure 11: Summary table with one row for every assessed object, segment...... 73
Figure 12: Administrative aggregation and higher-level statistics .................. 74
Figure 13: Limit-of-agreement graph for population estimates by two experts ...... 76
Figure 14: A geometric visualization of Bayes' theorem .................................... 92
Figure 15: Belief updating and weighted information gaps ................................ 97
Figure 16: Nepal - Information gaps over time .................................................. 98
Figure 17: A probability range updated in response to a test result .................... 100
Figure 18: Posterior probabilities in response to six cases of priors and test results.101
Figure 19: Process tracing tests and probable evidence ..................................... 105
Figure 20: A Bayesian belief network ................................................................. 110
Figure 21: Belief network about migration patterns in an area in Mali ........... 111
Figure 22: Definition of four scenarios ............................................................... 111
Figure 23: Example of a conditional probability table, Mali sample ............... 112
Figure 24: Experts with differing causal assumptions ....................................... 113
Figure 25: The ACAPS Analysis Spectrum ....................................................... 116
Figure 26: The general structure of a problem tree representation ................. 127
Figure 27: Problem tree of the expected Ebola impacts ................................... 128
Figure 28: Modified problem tree, example demand on health services .......... 131
Figure 29: Changing evaluative criteria, by stages of the analytic process ...... 135
Figure 30: Information flows and products in the Nepal Assessment Unit ....... 137
Figure 31: Shifting priority areas from day to day ............................................ 139
Figure 32: The risk analysis likelihood scale ....................................................... 148
Figure 33: Visual scales of probability and impact .......................................... 161
Figure 34: Four diagrams relating to the triangular distribution ....................... 174
Figure 35: Screenshot of an MS Excel VBA module ......................................... 178
Figure 36: Sample bias card ............................................................................. 186
## Acronyms and abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACAPS</td>
<td>Assessment Capacities Project</td>
</tr>
<tr>
<td>ACU</td>
<td>Assessment Coordination Unit</td>
</tr>
<tr>
<td>AWG</td>
<td>Assessment Working Group</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative distribution function</td>
</tr>
<tr>
<td>CESCR</td>
<td>United Nations Committee on Economic, Social, and Cultural Rights</td>
</tr>
<tr>
<td>DFID</td>
<td>United Kingdom Department for International Development</td>
</tr>
<tr>
<td>ECHO</td>
<td>European Civil Protection and Humanitarian Aid Operations</td>
</tr>
<tr>
<td>ERW</td>
<td>Explosive Remnants of War</td>
</tr>
<tr>
<td>ESCWA</td>
<td>United Nations Economic and Social Committee for Western Africa</td>
</tr>
<tr>
<td>FAO</td>
<td>United Nations Food and Agricultural Organization</td>
</tr>
<tr>
<td>FEWS NET</td>
<td>Famine Early Warning Systems Network</td>
</tr>
<tr>
<td>FGD</td>
<td>Focus group discussion</td>
</tr>
<tr>
<td>FSNAU</td>
<td>Food Security and Nutrition Analysis Unit</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>HPN</td>
<td>Humanitarian Practice Network</td>
</tr>
<tr>
<td>HRW</td>
<td>Human Rights Watch</td>
</tr>
<tr>
<td>ICDF</td>
<td>Inverse cumulative distribution function</td>
</tr>
<tr>
<td>ICRC</td>
<td>International Committee of the Red Cross</td>
</tr>
<tr>
<td>IDP</td>
<td>Internally displaced person</td>
</tr>
<tr>
<td>INGO</td>
<td>International non-governmental organization</td>
</tr>
<tr>
<td>IPC</td>
<td>Integrated Phase Classification</td>
</tr>
<tr>
<td>MSF</td>
<td>Médecins Sans Frontières; Doctors Without Borders</td>
</tr>
<tr>
<td>MSNA</td>
<td>Multi-Sector Needs Assessment</td>
</tr>
<tr>
<td>NFI</td>
<td>Non-food items</td>
</tr>
<tr>
<td>NGO</td>
<td>Non-governmental organization</td>
</tr>
<tr>
<td>NIF</td>
<td>Needs Identification Framework</td>
</tr>
<tr>
<td>NPM</td>
<td>Needs and Population Monitoring Project</td>
</tr>
<tr>
<td>OCHA</td>
<td>United Nations Office for the Coordination of Humanitarian Affairs</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>ODI</td>
<td>Overseas Development Institute</td>
</tr>
<tr>
<td>OSOCC</td>
<td>On-site Operations Coordination Centre</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability density function</td>
</tr>
<tr>
<td>PIN</td>
<td>People in need</td>
</tr>
<tr>
<td>RAF</td>
<td>Response Analysis Framework</td>
</tr>
<tr>
<td>RCRC</td>
<td>Red Cross Red Crescent Movement</td>
</tr>
<tr>
<td>SINA</td>
<td>Syria Integrated Needs Assessment</td>
</tr>
<tr>
<td>SNAP</td>
<td>Syria Needs Analysis Project</td>
</tr>
<tr>
<td>UNCA</td>
<td>United Nations Correspondents Association</td>
</tr>
<tr>
<td>UNDAC</td>
<td>United Nations Disaster Assessment and Coordination</td>
</tr>
<tr>
<td>UNDP</td>
<td>United Nations Development Programme</td>
</tr>
<tr>
<td>UNHCR</td>
<td>United Nations High Commissioner for Refugees</td>
</tr>
<tr>
<td>USAID</td>
<td>United States Agency for International Development</td>
</tr>
<tr>
<td>UXO</td>
<td>Unexploded ordnance</td>
</tr>
<tr>
<td>WASH</td>
<td>Water, sanitation and hygiene</td>
</tr>
<tr>
<td>WFP</td>
<td>World Food Programme</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organisation</td>
</tr>
<tr>
<td>WoSA</td>
<td>Whole of Syria Approach</td>
</tr>
</tbody>
</table>

In the main text, the first occurrence of an organizational acronym is preceded by the full name unless it occurs only in a literature citation.

Throughout the document, frequent reference is made to the roles of ‘expert’, ‘analyst’ and ‘decision-maker’. To avoid clumsy repetition of these nouns, personal and possessive pronouns in the singular refer to the ‘expert’ as female, to the ‘analyst’ as male and to the ‘decision-maker’ as he/she, his/her, etc.
Acknowledgments

We thank the following persons for their support:

  Agnès Dhur, Coordinated Assessment Support Section, United Nations Office for the Coordination of Humanitarian Affairs (OCHA), Geneva, deepened our understanding of humanitarian expert judgment in stimulating conversations.

  Lukas Drees, Water Resources and Land Use, Institute for Social-Ecological Research, Hamburg, made visuals available, and commented on our initial summary, of Bayesian belief networks in a humanitarian application.

  Anneli Eriksson and Johan Von Schreeb, Centre for Research on Health Care in Disasters, Karolinska Institute, Stockholm, helped with advice and literature.

  Jane Linekar, Ane Utne and Ana Escaso, ACAPS Geneva, commented on the analytic workflows within ACAPS, provided editorial help and improved the document design.

Indirectly, but no less importantly, this study benefited from comments on expert judgment as well as from the practical example by uncounted humanitarian workers and colleagues that we, the authors of this study, met in the programs and places where ACAPS was active.
1. Summary

1.1. About this study

Experts are indispensable in modern organizations. They fill gaps in data and in the understanding of existing or missing data. They introduce, apply and teach techniques and methods, some of which staff of the experts’ principals - and ultimately others - will continue to employ and disseminate. Technical experts reduce uncertainty by working out consensus opinions and probability ranges. Policy experts unravel the preferences and capacities of stakeholders; by doing so they dampen excessive certainty and may increase uncertainty in strategic ways that decision-makers and analysts find productive. When experts give their opinions in a context of decision-making, these become expert judgments.

The functional contributions of experts – data, interpretation, methods – and the professional roles that produce, process and consume judgment – expert, analyst, and decision-maker – are generic and universal. They pattern the insertion and work of experts in the humanitarian sphere, too. Yet, there are some important differences vis-à-vis other institutional arenas. Typically, the environment of humanitarian action is more turbulent than those in which expert judgment methodologies have matured, such as nuclear power plant engineering. This turbulence blurs the distinctions among experts and other roles, including decision-makers, analysts and key informants. It may also explain why there has been little systematic work about expert judgment methodologies for the humanitarian domain.

This study speaks primarily to humanitarian analysts. Analysts mediate between decision-makers and experts. They facilitate the ongoing expert work, aggregate the contributions of several experts, and edit the aggregated product in the perspective and dialect to which the organization and its partners are habituated. The goal of the study is to enhance analyst competence in dealing with experts and expertise. It offers insight also to decision-makers who commission and consume expert work, and to experts for whom some aspects of expert judgment theory may be new. Further, it may help stakeholders and academics situate the particular dynamics of expert judgment in the humanitarian environment.

We do not provide a comprehensive theory of humanitarian decision-making. Instead, we discuss tools and tasks that humanitarian analysts perform or oversee in the production of expert judgment. Our major focus is on relationships among decision-makers, analysts and experts, on technical aspects of eliciting and aggregating judgments, and on demonstrating, through case studies, how reality affects even the best-laid plans to elicit and use expert judgment.

We look at the work of experts who answer specific questions at a given point in time or within a few weeks or months at most in the context of humanitarian decision making. We rarely refer to long-running communities of practice organized around a broader topic, such as the Cash Learning Partnership or the Digital Humanitarian Network. For

---

1 Betts and Bloom, in their “Humanitarian innovation” study (2014), analyze those and other long-run innovation streams.
the wider context, we provide a snapshot of the evolution of modern expert judgment and discuss differences between humanitarian experts and others.

Figure 1: The process of expert judgment

The major chapters are devoted to the process of using expert judgment and to select applications. The process typically is described in stages that include background and preparation; the recruitment of experts; collection (“elicitation”) of experts’ opinions; the combination of their contributions through quantitative aggregation and qualitative synthesis; communication of the findings to decision-makers and stakeholders, and their actual use in decision-making.

The sections on background, preparation and recruitment are short. They emphasize the important distinction between the goals of the expertise, the broad question areas the experts need to cover, and the specific questions within the areas, as well as key decisions that decision-makers and analysts make before hiring relevant experts.

The elicitation and aggregation/synthesis chapters proceed in greater detail. They alternate between prescriptive guidelines and sidebars that illustrate past experiences, special methods and new developments. We dispense with the final stages on communicating expert judgments and their use in decision-making, not having enough experiential material to discuss these stages. Sufficient guidance can be found in the rich literature on communicating research findings.

The process chapters are followed by case studies of actual humanitarian applications, most of them drawn from ACAPS engagements. The final chapter reflects on history, the present, and future possibilities for humanitarian expert judgment. Expert judgment remains a resource that helps decision-makers find out and make sense of ‘what’s going on’, and determine what their organizations can and should do in response.

1.2. Background, preparation, recruitment

Organizations feel the need to involve experts most keenly when they fail to meet objectives or when important processes are deficient. Other conditions that call for expertise are informational. In particular, experts are more likely to be brought in when:

- normal data (e.g., from surveys) are sparse or flawed
- uncertainty is high
- experts are better, faster, or cheaper than other potential solutions
- new information requires frequent updates of assumptions and decisions
- additional validation is required
- available data are rich, yet some key parameters cannot be estimated
The situations that prompt decision-makers to hire experts are thus highly variable. Good preparation includes careful definitions at three levels: the overall goal of the experts’ work, the broad question areas they are supposed to cover, and the specific questions they must answer in order to inform decisions meaningfully. The questions must be such that experts who can answer them can be effectively borrowed or hired from existing networks and markets.

Other key parameters to be considered prior to recruitment are:

- whether the expertise is meant to be reassuring or strategically disruptive
- whether the objective is to obtain the experts’ answers or to learn their problem-solving processes
- the complexity of the areas and the questions, and the magnitude of the data required
- the number and types of experts needed, and whether they need to be recruited simultaneously, as opposed to sequential reviews and decisions to hire more or not
- the expected benefits weighed against the likely financial, opportunity and social costs of the expertise

These questions are resolved through an iterative process of refinement. This involves the sponsors (agencies funding the exercise, decision-makers using the results, workers interested to acquire skills from experts), the concerned personnel (managers, analysts, data collectors) and to variable degrees, the experts themselves. It is not unusual for sponsors to ask experts to write their own terms of reference, or to begin by hiring an expert tasked with designing a multi-expert program. The degree to which prospective or confirmed experts are to work out question areas and specific questions is itself an important parameter to be determined in the preparations.

Those are all generic requirements. What is specific to humanitarian experts? In popular notions, expert authority rests on superior technical and subject matter knowledge, acquired in long years of training and experience. However, humanitarian experts function in a multi-disciplinary milieu. Most of them are valued because of particular personal mixtures of subject matter, technical, geographic, language, social and cultural intelligence. They are less clearly demarcated from other positions than experts in other domains; the boundaries with agency staff, key informants, citizen scientists and media workers are fluid and often permeated by issues of access and availability. Humanitarian experts are less often tied to any of the classic professions such as medicine, law and engineering; their knowledge is more syncretic – coherent if drawn from motley sources.

As a result, humanitarian experts are experts because in the eyes of their principals, the public and other experts, they know something that is worth knowing about humanitarian needs, risks and response.

This self-referential definition is as inevitable as it is unsatisfactory. It is inevitable because the humanitarian work environment challenges professional boundaries. It is unsatisfactory because it may excuse vicious cycles of low expectations. In emergency settings, multi-sectoral gatherings and processes highly depend on agency
representatives with a certain level of expertise. Notoriously, agencies often send participants whose major qualification is that they are the most expendable on that day.

The multi-disciplinary milieu and the thin boundaries with other social roles have consequences for the recruitment of experts. The decision-maker decides whether the experts are primarily to reassure or to irritate, and for what purpose. The recruiters then look for an appropriate and manageable diversity of backgrounds, competencies and characters. Agency workers contribute much of the needed expertise, because they understand both the institutional agendas and relevant segments of the environment. Their use is more expedient than bringing in outsiders at additional expense and delay. The importance of local knowledge makes key informants indispensable experts.

Decision-makers and analysts must then narrow down the remaining gaps with the help of external experts. They need to find experts with the right mixture of substantive and formal knowledge, at the right level of granularity. They need to decide if the primary aim is to harness certified knowledge or rather to exploit institutional affiliations. On top of that, they need to square all the desirables with budgets for consultants, agency calendars and the brittle consensus of colleagues and stakeholders.

1.3. Elicitation and recording

“Eliciting” is the technical term for the analyst’s activity that causes the expert to form and express an opinion. It is a careful and carefully designed activity, down to fine detail; it adheres to “specially designed methods of verbal and written communication”.

In a classic of the expert judgment literature, Meyer and Booker (1991:100 sqq.), from which we have taken the above quote, divide elicitation into five components:

**Elicitation situations** comprise one-on-one interviews between interviewer (analyst or data collector) and the expert, interactive expert groups, and Delphi techniques in which experts see the judgments of other experts indirectly.

**Mode of communication**, such as face-to-face, telephone and computer-aided (chiefly Web-based).

**Elicitation techniques** range from the time-consuming ethnographic approach (where the analyst rephrases the expert’s responses continuously into new questions), to verbal reports (the expert thinks aloud as he works towards a solution; the analyst records) to less onerous verbal probes (the analyst asks questions only after the expert reports his solution).

**Response modes** are the formats in which the experts are asked to encode their judgments, such as probabilities, ranks or ratings.

**Aggregation** is the combination of several pieces of information into one statement. This is part of the elicitation process only if it is done by the experts themselves. Aggregation is “behavioral” when it results from interaction among experts, as opposed to “mathematical” when it is the outcome of algorithms an individual expert employs.

The combinations and refinements are almost limitless -- the elicitation chapter in Meyer and Booker runs over a hundred pages. This amount of guidance cannot and
should not be absorbed until needed. Even then, one should proceed selectively and with firm resolution to adhere, and have data collectors and experts adhere, to the arrangements that are critical for the particular purpose.

Regardless of specifics, it is always helpful to evaluate elicitation options against two general insights. First, and unsurprisingly, experts are subject to the limitations of human information processing. The interaction between analyst and expert, and among experts, operates similarly to the interviewer-respondent dynamic in questionnaire surveys. Survey methodologists have broken it down to four constituent operations:

- Comprehension – respondent interprets the questions
- Retrieval – respondent recalls from memory the information needed to answer
- Judgment and estimation – respondent combines and summarizes the information they recall or substitute an estimate for a missing element
- Reporting – respondent formulates response and puts it in the required format (Groves, Fowler et al. 2004:202).

The elicitation format needs to respect the limitations of the expert at every stage, even when the work extends beyond a single encounter.

Second, the turbulence of the organizational environment limits the complexity viable in elicitation instruments. High expert turnover means more time spent on repeated briefings and loss of historic depth. Rapid changes in target populations entail higher estimation error in the updates that key informants provide. Regardless of whether the turbulence is driven from inside or outside the agency eliciting expert judgments, it limits what can be extracted and transacted. Turbulence exerts pressure for simplification – it places a premium on robustness and reliability at the expense of detail, precision, and sometimes validity.

On the analyst’s side, proper arrangements to record the experts’ judgments and, as much as desired, the operations that formed them must be considered in the very design of the elicitation. In addition to data and statistics, the observation bases of the individual experts need to be established and recorded. It is one thing to collect estimates about the proportion of destroyed buildings from four experts, and another to know that expert A rests her estimate on his visits to two villages since the earthquake, B has been to four, C to 15 and D to at least 20. This information is critical in the aggregation phase.

1.4. Aggregation and synthesis

After information has been collected from multiple experts and properly recorded by the analysts, the next step is to “aggregate” it. “Aggregation” has several meanings that need to be kept separate.

In quantitative contexts, an analyst reduces the estimates that the experts have produced of the same variable to one combined value. Ideally, the aggregation produces also a measure of confidence or concordance, such as the confidence interval around the estimate of a continuous variable or of a proportion. Aggregation in the quantitative context also refers to operations that combine counts and estimates made on low-level administrative or geographical units. They produce statistics of the next higher level(s),
chiefly by summation and weighted averages. Importantly, the uncertainty measures are not additive, but need to be recalculated at every level.

In the qualitative domain, “aggregation” is a misnomer. Propositions can be counted from their occurrences in expert reports, but such frequencies do not constitute a summary. When analysts summarize qualitative information from multiple experts, they need to find abstractions on a higher level, perhaps borrowed from academic theories. The process is more aptly called “synthesis”, the term we will use in this study.

We were, however, unable to find pertinent literature guiding the synthesis of qualitative expertise in the humanitarian world. We take guidance from outside, from Sandelowski and Barroso’s *Metasynthesis of Qualitative Findings* (2003). These authors lay out a solid, practical three-step process. In step one, the analyst collects, abstracts and orders the findings from all the contributions under review. In step two, he reorganizes the findings, comparing them in multiple ways. In step three, he extracts the experts’ own syntheses, imports external concepts and fuses interpretations in their light. By concentrating on this one approach, we cut a path through the sprawling thicket of qualitative methods.

The quantitative section of the chapter has two objectives. First, we present methods for some of the most common aggregation situations. This part is extensive; most readers will want to absorb certain parts, as and when needed. We present each method briefly with its defining situation, purpose, logic and caveats. We amplify most technicalities in sidebars. Where we provide worked examples, they are screenshots of Excel spreadsheets, with the less obvious formulas visualized.

We discuss four frequently encountered situations:

A. Scalars (real-valued unconstrained variables) when:
   1. Experts state their uncertainty vs.
   2. Experts give only point estimates, but the experts are the same for all objects

B. Proportions and probabilities when:
   3. The observation bases of the individual experts are known vs.
   4. The observation bases are not known

The second objective is to create a basic familiarity with Bayesian ways of thinking. Humanitarian information is frequently updated; Bayesian thinking is all about how beliefs change in the light of new evidence, such as that produced by experts. Modern statistics is in the grip of a Bayesian revolution. Many humanitarian analysts will work with new methods from this wellspring at some point, providing added incentive to learn Bayesian basics early on. This section is limited to the exposition of Bayes’ Rule and to an intuitive grappling with some conceptual notions and qualitative applications useful in the humanitarian context. It also discusses “process tracing” as a method to test assumptions about causes and effects, on the cusp between the qualitative and the quantitative.

Throughout the chapter, the importance of characterizing the uncertainty surrounding experts’ estimates and interpretations is a recurring issue and constant reminder of good
elicitation design. Analysts should strive, as much as possible, to have experts provide such measures.

1.5. **Humanitarian applications**

**ACAPS’ Analysis Spectrum**

The chapter “Humanitarian applications” provides ten case studies of expert judgment in practice. The cases are arranged in the order in which ACAPS defines the “spectrum of analysis”: starting at the exploration level, working its way up increasingly from individual to shared work, through description, explanation, interpretation, anticipation, and eventually prescription. All the cases are from ACAPS’ own work, except the one on prescriptive expert judgment. The decision to develop case studies from ACAPS’ experiences was primarily driven by the lack of other documented humanitarian examples – it does not mean to suggest that ACAPS is the only actor which uses expert judgement as an informal or formal data collection method. Brief “Lessons learned” pointers conclude each case study.

**Case studies**

This brief listing gives the titles, topic and motivation of each of the ten cases:

- **Exploratory analysis: Using experts to find information**: ACAPS’ “Refugees/Migrants in Europe” project ran from December 2015 to March 2016. This first case study illustrates ways of harnessing expertise to the collection of dynamic information when other methods would be unacceptably slow, expensive or inflexible.

- **Descriptive analysis: The population inside Syria, 2015**: ACAPS helped the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) regional office in Amman, Jordan, with the design and analysis of population estimates that key informants based in the 270 sub-districts of Syria communicated, mostly by mobile phone. The case demonstrates the use of probabilistic methods as well as the need for flexible improvisation.

- **Explanatory analysis: Problem tree Ebola**: A major lesson learned during the Ebola in West Africa 2013-16 epidemic has been the need to broaden the scope of the humanitarian response during a large-scale outbreak. The disruption of public and private services created an “emergency within the emergency”. ACAPS used a device known as “problem tree” to represent the complexity of needs beyond disease control. The ACAPS Ebola problem tree served as an advocacy document to approach donors and finance a specific project dedicated to the analysis of the Ebola crisis.

- **Interpretative analysis: Severity ratings in complex emergencies**: In the first half of 2013, groups of humanitarian responders conducted three needs assessments in northern Syria. ACAPS personnel were involved in all three. The case demonstrates the use of expert judgment in the development of assessment instruments, here specifically in the measurement of severity, and in the organization, processing and analysis of the sub-district-based field assessment data.

- **Interpretative analysis: Clarifying priorities**: In the early stages of needs assessments and response planning, priorities may change frequently. In this
regard, the updating mechanism used by the Nepal Assessment Unit is of special interest. The Unit was activated by the United Nations Disaster Assessment and Coordination (UNDAC) in Kathmandu days after the first of two earthquakes that struck Nepal in spring 2015. The case is a study in expert judgment under time pressure.

- **Interpretative analysis: Decision support in an information-poor environment:** In 2016, two ACAPS analysts were part of the UNDAC assessment team in Haiti following Hurricane Matthew. They made a running inventory of all the assessment information arriving at Port-au-Prince. The mapped assessment gaps were widely noted in government and humanitarian agencies and caused the assessment activity to ramp up. The work was based on a methodology that ACAPS had already developed elsewhere, and which its analysts adapted to the local conditions.

- **Anticipatory analysis: Qualitative risk analysis:** On a monthly basis, ACAPS produces analysis of particularly relevant or dynamic specific risks. The project relies primarily on secondary data, rather than on the opinions of external experts. A team of ten analysts review the secondary data, and in this case are responsible for providing the expert judgment. The major challenges include maintaining the consistency by which the analysts identify risks and estimate likelihood\(^2\) and impacts. Information does not always arrive in good time in order to recognize deteriorating situations early. The weekly production rhythm is so short that the analysts tend to focus on the present, at the expense of the broader picture.

- **Anticipatory analysis: Scenario building:** ACAPS has several times brought together experts in scenario-building exercises. It relies on expert judgment in order to discern possible clusters of developments that can occur in crisis areas. Past exercises addressed the armed conflict in Nigeria, food security in Indonesia, the European refugee crisis as well as developments in Syria. This case study primarily looks into the diversity of workshop and meeting formats and their respective suitability for the different stages of the scenario process.

- **Anticipatory analysis: Visualizing impact and probability:** Some of the classic expert judgment methodologists were opposed to assigning probabilities to scenarios. ACAPS lets workshop participants discuss and, if possible, agree on probabilities. This has proven challenging. The case discusses, reporting the experience of two workshops in Nigeria, key issues as well as ways of visualizing impact and probability for easier understanding.

- **Prescriptive analysis: Experts vs. decision-makers:** ACAPS doesn’t prescribe, except through methodological advice and demonstrated analysis. For a prescriptive-level case study, we turned to outside experience. The case – strategy change during the 2010-11 famine in Somalia – was selected because it demonstrates the often difficult relationship between expert judgment and decision-making.

\(^2\) In the major part of our text, we use the terms “likelihood” synonymously with “probability”. In statistics, they have different meanings. In the section on Bayesian analysis, we distinguish the two, without the trouble of a systematic treatment.
1.6. Conclusion

“Expert opinion”, in the words of a classic author in this field, “is cheap, plentiful, and virtually inexhaustible” (Cooke 1991:3). That is an exaggeration. At a minimum we can say there is an established science of expert judgment. Some of its tenets transfer to the humanitarian domain almost in their entirety. The imperative to clearly define expectations at three levels – overall goal, broad question areas, specific questions – is a good example.

Other aspects are significantly modified as we move from generic expert judgment methodology to applications in the humanitarian sphere. The humanitarian environment is turbulent. The turbulence has consequences for the functioning of expert judgment. Notably:

- The roles of decision-maker, analyst and expert are less differentiated than in traditional expert settings.
- Local knowledge is indispensable, making the key informant as important as the technician.
- The bases of decisions grow out of sequential contributions, rather than from the consensus of experts convened simultaneously.
- The collection and aggregation of judgments have to overcome language, access, conceptual and skill barriers.
- The high turnover of personnel stunts methodological traditions; expert judgment is not yet widely recognized as an established methodology.

That is what we have seen so far. Change is pervasive, also in expert judgment. It is a safe bet that the development of Web-based applications and the Big Data revolution will unsettle the ways experts are recruited and work in the humanitarian world. Pressures for “evidence-based policies” and “value for money” will be passed on to decision-makers, and from them to analysts and experts.

It is harder to predict specifically how this will happen. Technologies that generate and process massive data of quantitative and categorical nature may be adopted with relative ease. We may also see innovations in dealing with qualitative expert judgment, such as in Web- or Intranet-based “argumentative Delphi”, a technique that generates, evaluates and combines arguments and counter-arguments from a potentially large group of participants. Technique and infrastructure will connect novel coalitions of experts, key informants and even persons directly at risk in the evaluation of needs, environments and response options.

Humanitarian experts’ greatest contribution, however, does not depend on technological savvy. It is intellectual and emotional, helping decision-makers to confront “What’s the story here?” and “Now what shall we do?” questions. It is sense-making. The kind of sense-making that, as another classic put it, “involves turning circumstances into a situation that is comprehended explicitly in words and that serves as a springboard into action” (Weick, Sutcliffe et al. 2005). If this study encourages readers to better prepare for this task, it will make sense, indeed.
1.7. The way forward

The use of structured expert judgment within humanitarian decision making is far from common practice. This study encourages its wider application, through case studies that demonstrate its scope, as well as prescriptive chapters that provide solutions to specific problems. Above the flurry of historic and technical detail, our final paragraphs of the summary note some concerns that speak to the development of the wider business of humanitarian analysis, and the place of expert judgment within it. They take aim at longer-term attitudes, rather than immediate practices:

Shift the current focus on “data quality” to “conclusion quality”: Currently, the humanitarian community spends a lot of energy on improving data quality, and the debate continues on how best to achieve this. The focus on more and better data is incomplete, and sometimes misplaced. Perfectly valid conclusions can be derived from poor or insufficient data, and wrong ones from abundant and reliable data. If methods for assessing data quality have advanced, humanitarians have made scant progress in securing the quality of inference and interpretation. The wider adoption of analytical standards and expert judgment – both as a process and as an information source – can help to redress this imbalance.

Advocate for understanding and use: For expert judgment to turn into a mainstream method, those in a position to influence how data are generated must advocate for its use when appropriate. This includes creating an understanding of its strengths and limitations among those who commission and use humanitarian analysis. Guidance is specifically required to accompany critical humanitarian (funding) processes such as the Humanitarian Needs Overview and Humanitarian Response Plans, as well as population estimates. By the time the term “EEJ” (Eliciting Expert Judgment) has become as common as “FGD” (Focus Group Discussions), analysts’ job descriptions should include familiarity with the method and techniques as an essential requirement.

Invest in wider applicability: Promotion of expert judgment should go hand in hand with an expansion of the current tools and guidance. At the time of this research, there was no common foundation of expert judgment concepts and tools in the humanitarian community. Standardized tools and guidance appropriate to this sphere are needed, specifically regarding

- how to determine and report levels of uncertainty, and when
- which structured analytical techniques apply to humanitarian settings, e.g. collaborative and joint analysis processes
- aggregation of expert judgment, e.g. for generating population estimates
- expert recruitment techniques and criteria, including for joint analysis workshops.

Other disciplines, particularly the decision sciences, should be queried for inspiration and guidance.

Use structured design and implementation: The study provides a set of recommendations on the actual production of expert judgment. While these skills and techniques belong chiefly to analysts, they thrive in a milieu in which leaders show an appreciation for analytic discipline. Expert judgment needs a system that is structured, planned and capable of incorporating other relevant information as required. The highly
dynamic nature of a humanitarian crisis requires a stable process, with committed leaders, trained staff and rigorous methodology.

**Capture and reflect uncertainty:** A clear understanding of uncertainty is required for a successful elicitation process, and everyone involved plays a role in creating this comprehension. Experts are to be open about their limitations, information gaps and biases. Analysts must ensure that the uncertainty of estimates and inferences are recorded and communicated. Dissent among experts should be welcomed as a source of new insight and as corrective to false certainty; analysts should record opposing positions and their rationales, instead of papering over them with rash consensus findings. An acceptance of uncertainty, and an understanding of how to use this within decision making, is essential for all consumers of humanitarian analysis.

**Recognize constraints and develop relevant approaches:** The humanitarian environment makes it near-impossible to “calibrate” (assess on comparable track records or formal tests) individual experts and thus weight their opinions arithmetically. Staff turnover in emergencies is so high that it thwarts even the milder ambition of using the same experts across multiple occasions or different crises. To improve access to experts and hence the analysis of their contributions, new collaborative platforms are needed. To counteract loss of institutional memory and of coherence, arrangements are needed for analysts to extensively debrief experts and key informants. Severe limits exist on working with multiple experts simultaneously (although media like Skype have relaxed them somewhat); strategies and methods for sequential recruitment and elicitation as well as for cumulative analysis need to be further developed. The Bayesian philosophy of updating one’s beliefs in the light of new evidence should gradually filter into humanitarian analysis from disciplines where it is strongly established.

**Documentation of practices:** The product of the experts’ work – often but not always a report – should describe the essential processes followed, the analytical techniques applied, the information sources used and the levels of uncertainty determined. Without these pieces of information, the product is diminished or even worthless to anyone outside of the persons directly involved. Documentation on expert work should at minimum state the decision-context in which the expertise was sought, how the experts were recruited, their terms of reference, and how their judgments were elicited, subsequently aggregated and finally translated into options and recommendations for the decision-makers and stakeholders. Transparency on humanitarian expert judgment practices serves an additional purpose. In a setting with limited recorded practice, openness on tools and methods is essential to inspire and strengthen a budding movement that may, with good nurturing, grow into a normal practice – the practice of expert judgment.
introduction
2. Introduction

2.1. What this is about

This study discusses the scope and use of expert judgment in humanitarian analysis. Expert judgment is expert opinion given in the context of a decision. Analysis in humanitarian settings is the structured, transparent and controlled human process of transforming raw data into actionable insights. Expert opinion is data as well as context provided by persons with supposedly superior skill or knowledge when availability, quality, time and cost considerations rule out traditionally sourced data. We address the study to humanitarian workers who fill one of these roles and may have particular learning objectives:

- Analysts (such as Information Management Officers) who are contact points, facilitators or supervisors for experts and/or who design, accompany and process the experts’ work
- Decision-makers who commission and consume expert work
- Current or aspiring humanitarian experts eager to familiarize with aspects of expert judgment theory that may be new to them
- Academics studying the intersection of expert judgment and humanitarian response

The major focus is on relationships among decision-makers, analysts and experts, on technical aspects of eliciting and aggregating judgments, and on demonstrating, through case studies, how reality affects even the best-laid expert judgment designs.

There is no agreed typology of humanitarian decision-making situations. We, therefore, do not insist on the difference between expert judgment and expert opinion in this note. Our objective is to situate humanitarian expert judgment in the generic tradition of expert judgment, to find common traits as well as differences between the two and to detail essentials of the judgment process. In particular, two chapters are devoted to the ways experts are helped to produce good judgments – a process known as elicitation – and to methods of combining expert judgments, as aggregation of quantitative estimates or synthesis of qualitative reports. Thereafter we describe a number of humanitarian applications; ACAPS has been engaged in most of them. These case studies will not traverse the full panorama of expert judgment types in modern collective action, of which the humanitarian is a sector besides many others. But it will give pointers to a limited number of techniques and resources likely useful in humanitarian analysis for some time to come.

2.2. From Machiavelli to nuclear power

In order to properly situate expert judgment in the humanitarian world, a brief view of the evolution of this method in the wider society is helpful. Throughout documented history, human societies have relied on persons with special knowledge. Professions evolved to respond to threats of death and destruction – physicians in the face of illness and epidemics, lawyers to domesticate looming violence, priests to help transition souls between earth and heaven. Individual decision-makers supplemented their own wisdom with professional advice. Over time, as society grew increasingly self-reflective, advice about how best to seek and benefit from advice was collected, reviewed and transmitted.
Introduction - From Machiavelli to nuclear power

as “advice about advice”. Renaissance rulers, ever prone to making catastrophic political and military decisions, rose and fell with their choice of able advisors. The work of Machiavelli (1469 –1527) is emblematic for this kind of calculated supply of, and demand for, strategic advice.

With increasing social differentiation, norms and organizations arose to deal, not with one, but with groups of advisors – the collegium of physicians at the bedside of wealthy patients, the auditors for the far-flung operations of a joint stock company, the engineers designing railway lines safe for trains at every point. The behavior of these professions was largely self-regulated, within the confines of law and custom, in the tacit assumption that their work effectively reduced uncertainty in stable technical and institutional solutions.

The catastrophes of the twentieth century destroyed that certainty. The military planners of the German Empire knew – from repeated war games before 1914 – that a rapid advance across Belgium and the northern plains might run afoul of stretched supply lines and of the advantage that a dense railroad network gave the defending party. But they had no means of quantifying the risk, failed to effectively communicate it to the military and political leadership, and were not free to develop scenarios for the longer range, including any that envisioned the collapse of the Empire. Lacking those, they were pre-modern experts only.

One World War later, the elegant solution to the “German tank problem” (Wikipedia 2009) was celebrated as one of the “Great Moments in Statistics” (Grajalez, Magnello et al. 2013). The problem was to estimate the monthly production of a certain type of tank by the Nazi war machine from serial numbers imprinted in components of just two captured exemplars. For this note, we highlight it as an ideal-type constellation of decision-makers, analysts and experts – key roles played in modern expert judgment. In this case, the planners of the invasion in Normandy were the decision-makers; they revised their plans when the analysts – statisticians applying an appropriate formula – submitted their estimate. An important piece in their model came in the shape of the expected number of tires that could be made with a given mould before it had to be replaced – an expert judgment supplied by British tire manufacturers. The high stakes in the problem, its resolute quantification despite a paucity of data, the provision of a critical parameter from several outside sources of knowledge – they all foreshadowed modern expert judgment.

The nuclear age completed the transformation of the traditional advisor, operating by virtue of charisma and undoubted reputation, into the methodologically supervised professional – the modern expert. The annihilation of humankind became a possibility, an event of (variably!) low probability and extreme consequences. The judgments of experts who brooded over the risk of mutually assured destruction needed to be reined in so that madness and rationality could co-exist. Think tanks like the Rand Corporation rapidly expanded the canon for military and strategic expert judgment.

On a less apocalyptic scale, the systematic elaboration of the risks that nuclear power plants were posing propelled the further development of expert judgment methods. Roger Cooke’s “Experts in Uncertainty. Opinion and Subjective Probability in Science” is a classic in the field (Cooke 1991). The book is largely guided by the kinds of experts who know a lot about, say, the half-life of a certain type of pressure tank and can be
induced to share this knowledge *in isolation* from other confounding variables. Over the years two major nuclear-plant disasters, besides a number of lesser accidents and numerous risky situations, have been witnessed by shocked publics and rationalized by surprised experts. The experience has greatly contributed to the development of *technical* expert judgment methodologies as well as to insights into the limits of this type of knowledge.

### 2.3. Modern expertise and its discontents

To claim that war, the risk of war and the civilian fall-out of military technologies are the fathers of all modern expert judgment would be plainly incorrect and barely palatable for a discussion of its legitimate uses in the humanitarian world. Not only have other institutional realms – medical diagnostics comes to mind – contributed and clarified new tools, but the limitations and dangers of purely technical expertise themselves have given rise to *counter-movements*. Some of these shade into the humanitarian world.

First, there is a counter-movement inside expert judgment methodologies. *Technical* experts seek to achieve a high degree of consensus, typically producing estimates of probabilities or of other quantitative variables, with residual uncertainties ideally eliminated or greatly narrowed down. By contrast *policy* experts seek to incorporate “the views of the entire spectrum of ‘stakeholders’ and seek to communicate the spread of their opinions to the decision-maker”, and do so broadly (Cooke, op.cit., 12-13, with respect to Delphi methods, but generally valid for the technical vs. policy perspectives). In social movements challenging military, corporate or other dominions, policy experts typically enjoy greater prominence than technical ones. The two overlap, such as in the mathematical study of welfare and democratic participation. To illustrate, the so-called “Borda count” (Wikipedia 2011b) as a method to aggregate preferences in humanitarian needs assessments is a small import from the world of generic policy expertise.

Second, the worldview of technical expertise has been increasingly challenged by the necessity to incorporate *qualitative information* and, with ambiguous results, a flurry of *qualitative methods*. Planning theory deals with so-called “wicked problems”, some elements of which may resist fruitful quantitative description (Rittel and Webber 1973, for a reflection on the humanitarian realm, see: Tatham and Houghton 2011). A problem is “wicked”, for example, when those supposed to solve it are part of it. In the social sciences, interest has exploded in mixed methodologies that bring together qualitative and quantitative information at different junctures of the research process. They are popular, not necessarily because the results are always persuasive, but also because researchers claim the use of such methods as a preemptive defense. The aggregation of qualitative and quantitative expert judgments is an area that has seen progress (Moss and Edmonds 2005, Van Der Sluijs, Craye et al. 2005, Bradley, Dietrich et al. 2014), but the methodological demands can be daunting. They are daunting also to humanitarian practitioners.

Third, the *acceptance of experts* fluctuates. Policy expertise becomes politicized almost naturally and therefore is both acclaimed and opposed. Technical expertise, even when it appears neutral, is not so by itself. Its brief is defined by extrinsic forces; apolitical experts too are coopted into power structures; the organizations that commission and consume the expertise not only have interests, but dictate the ways of seeing things, down to perceptual and linguistic conventions. When experts are proven
Introduction - Expert, analyst, decision-maker

catastrophically wrong, such as in the design of the Fukushima power plants, the public concludes that they are either incompetent or complacent if not outright corrupt.

Experts defend themselves by demonstrating that, whatever their political and moral commitments, they adjust their judgments to changing evidence and strive to do so in consistent, measurable ways (Tetlock 2005). Other defenses have experts widening the circle of competence, through layman science endeavors, collective analysis at the grassroots or social movement advocacy. Multi-disciplinarity and personal street credibility may dispel the impression of egg-headedness and elitism, but sometimes have the opposite effect of diluting focus and importing biases from an earlier career. The macro-effect of all that is a kind of equilibrium between suspicion and a belief in the inevitability of depending on experts³.

Experts are not hired for their scientific objectivity, but because the individuals and institutions that recruit them recognize that they have a need for new knowledge, in order to survive in rapidly changing environments, and it is experts who help them acquire and adapt it (very pointedly, for the structural coupling between the scientific and the political systems: Luhmann 1997:785-786). The upshot for our subject – expert judgments in the humanitarian world - is that placidity or turbulence of organizational environments are more decisive for the consistency and productivity of experts than professional competence⁴.

These historic and sociological reflections set the background for the discussion of expert judgment in humanitarian action. However, we must, belatedly, first offer a generic description of experts and expert judgments.

2.4. Expert, analyst, decision-maker

Expert judgments are opinions that experts give in the context of a decision. Opinions divulged by “experts” on a TV show may, in the eyes of their peers, qualify as highly expert. But in this situation the experts are not giving input to a decision, and their opinions should not, at that moment, be counted as expert judgments. In the decision context, the decision-maker commissions one or several experts to provide their judgments on questions and in formats that someone has laid down for the experts – the decision-maker, analysts plying between him and the experts or, because of the ignorance of the former, the experts themselves. The three roles of decision-maker, analyst and expert are constitutive; yet one and the same person may play more than one, if not simultaneously, then consecutively. In fact, organizations thrive because of the mobility of experts; at the same time they are criticized for the blindsiding collusion

³ A notorious illustration can be found in the prosecution and later acquittal of seven Italian earthquake experts accused of “failing to adequately warn” the residents of L’Aquila, a city hit by a 5.9 earthquake resulting in over 300 deaths (spinning off a rich debate, e.g.: Alexander 2014).

⁴ The concept of turbulent environments (of organizations) goes back fifty years (Emery and Trist 1965). A modest literature has since accumulated (Radford 1978, Camillus and Datta 1991, Wooldridge and Wester 1991, Salyer 1995, Therrien 1995, Grant 2003, Liao, Welsch et al. 2003), most of which is in management science, not expert judgment (Salyer on nurses is an exception). For a rule of thumb, the environment of an organization is the more turbulent the more frequent changes it produces, the larger the individual change, the more highly correlated different types of change. We do not know of any specific studies of the turbulence that the environment thrusts on humanitarian agencies, but it is plausible that disasters, violence and crises cause a level of turbulence more dramatic than what organization outside such conditions suffer. Firms in rapid technological change may rank next in this regard.
Introduction - Humanitarian experts

between self-interested managers and venial experts. This makes it difficult to define experts in a clean and neat manner.

The Wikipedia article on experts offers two competing definitions:

- An expert is somebody who obtains results that are superior to those obtained by the majority of the population.
- Alternatively, an expert is someone widely recognized as a reliable source of technique or skill whose faculty for judging or deciding rightly, justly, or wisely is accorded authority and status by peers or the public in a specific well-distinguished domain (Wikipedia 2016e).

The first understanding may be appropriate in areas in which knowledge and personal ability to perform coincide. One thinks of purely intellectual, rule-based games that form closed systems. There is no reason why the grand masters in chess and Go should not be considered experts; they play their game better than most others and can fairly well predict the outcomes of other people’s games looking at current positions. For open systems, this is less often true: experts are not systematically better than laypersons at predicting outcomes in their fields (Camerer and Johnson 1997, for an experiment, using humanitarian experts as well as laypersons, see Parker 2017: chapter 4). This understanding of what makes an expert is not productive for our purpose.

The second reflects a popular consensus. It may exaggerate the reliability of experts; the “faculty for judging or deciding rightly” of experts is not invariably better than that of laypersons, let alone of mid-level technicians in the same fields. It may be fairer to say that experts know what one needs to know in their fields, and know where to find the knowledge. They know how to update it and, with the right incentives and directives, how to share it constructively.

In the core institutions of modern society, the knowledge that needs to be known in a particular field is organized around professions (Abbott 1988). Professionals are distinct from technicians by virtue of abstract knowledge that, in each classic profession, they command in far-reaching autonomy from other professions – a physician may need to consider the risk of litigation that would expose her to lawyers and judges, but her diagnostic toolbox is entirely medical. Abstract conceptual knowledge is the hallmark and currency also of experts, professionals used in decision contexts.

2.5. Humanitarian experts

Are they experts?

It is obvious that the link between expertise and the organization of knowledge in the shape of professions poses a challenge for the definition of experts in the humanitarian field. Humanitarian action thrives with the productive collaboration of multiple professions, as well as of persons whose individual background is multi-disciplinary and, for some, chiefly experience-based. Thus, by the strict standard of autonomous abstract knowledge, it is non-professional. The humanitarian expert, therefore, is difficult to demarcate from the experienced technician or the local key informant.

---

5 This is the premise of the entire “participatory methodologies” movement (for many: Chambers 2008).
A missionary, trained as a theologian and working in some quiet corner of Africa for thirty years is not by profession a humanitarian expert. If, suddenly at the outbreak of a crisis, humanitarian agencies seek her input on the ambient local society, the missionary is more than just a key informant. She is an expert on account of her privileged perspective as an educated stranger, her command of the local language and of the concepts that it vehicles, as well as of this long exposure to elements of the situation that the agencies seek to comprehend. In other words, experts of the humanitarian domain cannot all be tagged to classic professions and careers. The expertise harnessed to the response admits more of a situational element. Local language, culture, social capital and history matter on a par with the experience, technical skill and conceptual armor of the agency worker.

The example of the old-hand missionary, far-fetched as it is, illuminates the open boundary between professional and local knowledge that humanitarian agencies bring together. To discuss expert judgment in the humanitarian domain in more general terms, we construct an equally artificial, but more abstract example, from which we then shall branch out to enumerate its various functions. With a bit of background, we present the example as a dialogue between a decision-maker and an expert, who is also the analyst.

A motivating example

In a region visited by a dangerous epidemic, a medical coordinator has been responsible for an area comprising some one hundred villages. For the last twenty days, no new cases have been admitted to the treatment center in the district headquarters. The coordinator has come under pressure to close the center once the last patients can be discharged and to transfer its personnel to one of the areas still in acute need.

To carefully assess the risk of new outbreaks, the coordinator in the past week had tracking teams visit all households in a sample of twenty villages. The reports came in yesterday; all were negative. The proportion of surveyed villages with new cases thus is zero.

That is encouraging, but must not be taken to mean that the risk of new outbreaks is zero. It is not uncommon even for a substantial risk (given the seriousness of the disease) to have all negatives in a small sample. Thus, if the risk for villages to see new cases is an evenly distributed 5 percent, the probability of finding zero positives (zero villages with cases) in a sample of 20 is 

\[(1 - 0.05)^{20} = 0.95^{20} = 0.358\] or 36 percent.

Note that this discussion is not about the professionalization of humanitarian workers, which has progressed greatly, particularly with the impulse of the 1994 Rwanda crisis, in specialized training institutions and curricula, conferences and behavior codes. Hugo Slim offers useful reflections on this dynamic (Slim 2005). His argument is essentially value-based – ideals, principles and behavior codes – and much less into the abstraction and autonomy of bodies of knowledge. More recently, James offers a cautious assessment of humanitarians as professionals (James 2016).

The story is a free adaptation, with our own numeric assumptions, of a motivating example for Bayesian analysis in the Stata manual, version 14, chapter on Bayesian analysis (Stata Corporation 2015:2-3). The authors credit Hoff (2009:3) with its original form. For now, we omit the Bayesian terminology. For an Internet-accessible article that discusses the pooling of data in the context of estimating the rate of rare events such as the ones in this story, Young-Xu and Chan (2008) may be helpful.
For the coordinator, therefore, the sample survey results are not enough to inform his decision. He determines that if he can find 1. experts to quantify the probability of new outbreaks from similar situations, 2. combine their estimate with his new evidence, and 3. thereby infer an expected proportion of villages with new outbreaks < 0.02 (< 2 percent), then he can responsibly close the center and move resources where they are needed more importantly.

An epidemiologist comes forward to say that, drawing on resurgent outbreaks in a number of similar epidemics, the experience was that between 1 and 6 percent of village communities would see new cases after three weeks of quiet, with an average of about 3 percent. She shows this distribution in the blue line of the diagram below. She adds a second curve that distributes the probability taking into account the recent survey of 20 villages. She informs the coordinator that in the light of this new evidence he should expect 2.3 percent of villages developing new cases. This is higher than his cut-off point. She advises him to survey another 20 villages. If all of these turned out negative, the risk would be updated to 1.9 percent of villages only, i.e. to a level that justified closing the center, as seen in the third curve.

The curious medical coordinator asks the epidemiologist how she arrived at those results. She explains that the experience of previous epidemics motivated her to model the probability of a village developing new cases. She relied on the so-called beta distribution, which is described by two parameters, alpha and beta. Evaluating the experience, she set $\alpha = 2$ and $\beta = 64.667$. “From theory”, she goes on, “we know that, under these assumptions, the expected proportion of villages with new cases after three weeks is $\frac{\alpha}{\alpha + \beta} = \frac{2}{2 + 64.667} = 0.03$. That is our best estimate.
And that is \textit{before} I included your survey finding. Taking into account that you had \( n = 20 \) villages surveyed, and \( m = 0 \) turned out to have new cases, I used this information in the updating formula

\[
\text{new\_proportion} = \frac{\alpha + m}{\alpha + \beta + n - m} = \frac{2 + 0}{2 + 64.667 + 20 - 0} = \frac{2}{86.667} = 0.023 > 0.020.
\]

The estimate is still higher than the cut-off point that you set for your decision to close the center, although not by very much. However, if you survey a random sample of 20 more villages, and no positives are found, then the expected proportion drops to

\[
\frac{\alpha + m}{\alpha + \beta + n - m} = \frac{2 + 0 + 0}{2 + 64.667 + 20 + 20 - 0 - 0} = \frac{2}{106.667} = 0.019 < 0.020.
\]

But remember, the proportion of 1.9 percent is just a best estimate\(^8\); there is considerable uncertainty around it – the underlying epidemic mechanism could be stronger or weaker. And, whatever it is in reality, your cut-off point is arbitrary, inspired by the ethics of sharing humanitarian resources. Your decision is the outcome of both observations and preferences\(^*\).

"Thank you", says our coordinator, "for your detailed, if slightly challenging explanation. I won’t doubt a bit of it; I trust the expert. But one thing confuses me: If I understand you correctly, you are giving me a probability \textit{distribution} of a proportion. But my seniors and colleagues want a simple probability, just \textit{one number}, for all the hundred villages to remain free of this terrible scourge."

She quickly types figures into a calculator and tells him: "You can do this for yourself, assuming that, with the 20 surveyed villages all negative, the expected proportion of villages with new cases is 2.3 percent. Thus the probability for a randomly picked village to remain free is 97.7 percent. For all 100 to remain free, assuming these are independent events, the chances are \( 0.977^{100} = 0.098 \), only a ten percent chance. But even with 40 surveyed villages all negative, your desired outcome has a chance of \( (1 - 0.019)^{40} = 0.981^{40} = 0.147 \) or 15 percent. At any rate, you are betting against the devil.

However, you may still be right. Perhaps the probability distribution that I estimated from previous epidemics is not relevant for your situation. Given the high-quality surveillance and case management that your team maintained throughout the past months, the historic range of 1 to 6 percent of communities with new cases is too pessimistic. Under your leadership there may have been learning effects that my model has not yet incorporated. Expertise suffers from its own obsolescence unless we find the means to regularly update it. After a few more months, the experience of your area will become a valuable new addition to the basis of our so-called expert judgment. Meanwhile, if your seniors want absolute certainty, you have no choice but to maintain your surveillance of every one of the hundred villages."

---

\textbf{Yes, they are experts}

What general points can be puzzled out of this fictitious episode?

\begin{itemize}
  \item There are situations in which decision-makers do obtain relevant, valid and reliable data, but the data by themselves do not yield the kinds of estimates that
\end{itemize}

\textsuperscript{8} For simplicity, we make no finite population correction.
would inform the decision to be made. The additional judgment of experts is needed in order to build bridges between the data and some of the decision criteria.

- The expertise itself is uncertain. Its purpose is to transform false certainty and misplaced or exaggerated uncertainty into reasoned and measured uncertainty. The uncertain, yet expert judgment makes the risk bearable for the decision-maker.
- Expertise, like other knowledge, over time grows outdated, obsolete and even dangerous unless it is regularly updated. In theory, the outcome of every application of expertise could be used to update it; in practice, learning from experience is limited by observational, coordination and orthodoxy costs.

Starting from there, we can enumerate a number of typical situations that make expert judgment beneficial. These situations occur in humanitarian as well as in many other institutional realms. Because humanitarian workers routinely deal with them, it is reasonable to assume that there are humanitarian experts:

**When “normal data” are sparse or flawed:** Low administrative capacity, whether endemic or caused by disasters and crises, reduces the flow of statistical data. The data that still come forth may be defective – less precise, more biased, incomplete, inconsistent or useless for the kinds of decisions at hand. Experts may fill the gaps with context and specific estimates based on long experience and recent exposure. Yet the adequacy of their conceptual framework and the quality of their most recent information deserve scrutiny.

Starting in 1999, the Global Landmine Survey produced nearly complete inventories of communities affected by Explosive Remnants of War (ERW) in a dozen or so countries. Country surveys relied on one-day participatory group interviews and minimal mapping of affected communities. Individual communities were assigned one of three priority categories based on a composite measure of recent victims, resources blocked and types of explosives. However, in order to validate predictions of the risk of mine strikes, the relationships among indicators were investigated through statistical models that looked at all surveyed communities in a country at once.

A comparison of findings from Yemen, Chad and Thailand revealed that community population size as well as several indicators of conflict intensity had effects of comparable significance on the risk (Benini, Moulton et al. 2002). To this extent, the expert judgments were successful. The survey designers failed to create comparable measures of the communities’ institutional endowments and their effects on risks. Communities with zero recent victims typically were: “technically more modernized” (Yemen), with “stronger civil society” groups (Chad), with “more diverse financial services” (Thailand) – each of them significant, but together not comparable.

**When uncertainty is high:** The data may be rich and reliable, but the decision-maker is uncertain about context and about the right model that connects the data to the decision criteria. The expert details a cause-effect scheme for the problem at hand,

---

9 The obsolescence of expertise may be accelerating worldwide. “The half-life of knowledge is shrinking” is an often quoted adage of post-modernism. But learning techniques are also evolving, even for experts.
demonstrates the extent to which the data support estimates of critical relationships, and
provides context for data and decision.

In the 1990s Operation Lifeline Sudan emergency food deliveries were severely
hampered by the unpredictability of flight clearances. The program reacted “by
developing more and more access locations and by applying for flight clearance to more
locations than its monthly work programs would require on a strictly technical basis”
(Benini 1997:343). Its experts discovered a synergy between the clearance patterns and
waves of bush airstrip preparations that local commanders and chiefs supported in order
to obtain relief: more feasible locations – more clearance requests – more of them
approved.

**When experts are better, faster or cheaper:** Collecting data supposed to inform a
decision may be foreseeable less effective or less efficient than accessing expert
judgment. If the decision premises are known except for those that experts can readily
fill, additional data collection for the immediate decision is unnecessary. As flood
waters recede, the question whether a particular relief crop should be promoted at this
time of the farming calendar may be settled with phone calls to a small number of
agronomists and seed suppliers.

**When new information requires frequent updates of conclusions / decisions:**
Decision-makers need models that keep them one step ahead of rapidly changing
situations, about which new information keeps arriving, sometimes in less than
satisfying patterns. Forecasts result from the application of domain-specific models that
experts build and evaluate. Global (Hufnagel, Brockmann et al. 2004) and local (Ristic
and Dawson 2016) epidemic forecasts are cases in point.

**When additional validation is required:** The meaning of “additional validation”
differs, depending on whether we seek input from policy experts or from the technical
kind although there are overlapping areas:

- An **individual decision proposal** may be run by policy experts, who give opinions
  on the likely acceptance by, and consequences for, the sets of stakeholders
  whose capacities and preferences they understand.
- When **multiple possible situations and interventions** are to be considered,
  groups of policy and technical experts can build and evaluate select scenarios.
- For thousands or **millions of possibilities**, simulation models may be built by
  statistical experts, who then present samples of hypothetical outcomes to
  substantive experts for plausibility checks (e.g. a set of simulated wild fire area
  maps handed to regional foresters; Thompson, Haas et al. 2015).

**When data are rich, but some model parameters cannot be estimated from them:**
The model and the data to calculate the quantities of interest may all be available, except
for some critical parameter values, on which expert opinion is canvassed. Typically, the
situation occurs with detailed budget plans for the finalization of which certain prices
(unit costs) are not known from the data. An example might be the cost function for
small-aircraft relief goods transports in South Sudan. Tenders come from suppliers who
do not reveal their costs, but aviation experts may be able to supply credible price ranges
for the negotiating customer.
The problem is not limited to quantitative models. In the reconstruction phase an education ministry may have plans based on detailed school assessments that specify the proposed supply of education. The demand for education may be unknown as regards its sensitivity to school fees and to the opportunity cost of children not working; child protection experts may help create policies and incentives that take the underlying issues into account, in a qualitative language.

### 2.6. Differences and commonalities

**Between humanitarian experts and non-experts**

Cooke (see above, page 9) found the distinction between “technical” and “policy” sufficient for an expert classification. An expert was either one or the other kind. Cooke’s approach to expert judgment was driven from the technical side. Policy experts were needed on purely technical grounds: the preferences of multiple decision-makers could not be rationally aggregated (Arrow's Impossibility Theorem: Wikipedia 2013). Policy experts would evaluate and arrange approximate procedures to sort and rank alternatives and foster consensus.

Yet, to work effectively, such experts too needed a relatively stable, placid environment. Unfortunately, the humanitarian environment is mostly not placid; it is in various states of turbulence. The purity of Cooke’s distinction breaks down.

In a typical UN or humanitarian INGO field office, professionals with more technical expertise work side by side with persons who are, to various degrees, policy experts. Although an agency or particular teams in it may pursue a clear substantive focus, turbulent environments favor multi-disciplinary competences. These are present in team composition and often also in the career background of a particular staff member. Both technical and policy experts may have competences that emphasize the geographic region, particular languages or the social and cultural context.

Beyond that, common characteristics of humanitarian experts are not easy to distil. The boundary between experts and other knowledgeable people is fluid and situational. Here is a list of groups and what plausibly sets them apart from experts:

- **Key informants:** The common assumption is that key informants contribute superior local knowledge whereas the experts command more abstract conceptual systems to evaluate and transform it, mixed with external and generic knowledge. However, key informants like our hypothetical missionary in Africa (see above), sought as an interpreter of the local culture, undermine that hierarchy. In International Committee of the Red Cross (ICRC) delegations in former Soviet republics, the field officers often were more highly educated than the delegates under whom they worked. While they did not enjoy the same status and lacked agency-specific frames and connections, they effectively supplied important elements of the humanitarian intelligence that laid out decision alternatives. The boundary between experts and key informants is fluid; some key informants, as the field officer example shows, are recruited into the humanitarian agencies. Some become experts serving a succession of individuals, positions and even organizations.
• **Lay persons and focus group participants**: A similar argument can be made about the difference between experts and focus groups composed of lay persons. The difference in this case may be greater because focus groups typically are short-lived. Their members’ educational backgrounds may be more modest, such that the same level of dialogue as between expatriate and field officer cannot easily be achieved. Yet, not to forget, focus groups are called because they bring together people who are, or were during the height of the crisis, daily struggling to cope, even to nakedly survive. This gives them authority to speak about needs, problems and solutions in ways that put them on an equal footing with experts, in terms of relevance and validity, if not of technical fluency.

• **Lay persons and “citizen science”**: The involvement of ordinary citizens in scientific endeavors has a long tradition when most scientists were unpaid, earning their lives in other occupations, or were independently rich (Silvertown 2009). However, the concept of “citizen science” packs a large number of activities, only some of which overlap:
  The participatory development movement has long built a case for “citizen science”. Lay persons, particularly among populations affected by expert-supported policies, have science-like knowledge that should be incorporated into the expertise. As an often given example, poor farmers have been known to make controlled experiments with seed varieties. While some of that serves the defence of traditional cultures, crowd-sourcing initiatives are enriching citizen science from the latest technological edge (Wikipedia 2016c). They have found their way into disaster management, as data gathering (e.g., mobile phone data to track population movements. A case from Haiti is described in Bengtsson, Lu et al. 2011) or as collaborative computation (such as in GIS support for map making). While the former holds great potential due to its real-time aggregation, it is only the latter that deserves the attribute "citizen science". A natural tension exists between these scattered contributors and the experts advising on the encompassing systems and policies. It revolves around data validation, scientific standards, and feedback as a reward for participation and for mutual improvement. Crowd-sourced expert opinion approaches are being developed (Bakdash and Marusich 2015) – but, then, how do we vouch for the experts in the crowd?

• **Academics**: A large proportion of experts are academics. Universities hire them because they already are recognized experts; and others have become experts because their universities and academic networks propel them into expert positions. Academics bring views into play that are empirically more broadly based or better secured in encompassing theories. Conversely, expert jobs demonstrate organizational affiliations and inspire studies that in turn help academic careers. In that sense, the boundary is highly porous; and it virtually disappears in academic institutions set up specifically for research into humanitarian affairs. Nothing of that is new or disturbing. Issues arise when client expectations compromise scientific integrity, or, conversely, when
Introduction - Differences and commonalities

academic culture holds the expert captive to perspectives and language that frustrate comprehensible judgment and pragmatic advice.

- **Media workers:** There is no reason why media workers steeped in areas and problems of humanitarian interest should be denied the title of expert. A journalist who has covered a country in conflict for many years, cultivating intensive contacts with relief and protection agencies, has valuable expertise. Yet the expertise of media workers is different from the technical and policy varieties. Technical experts are bent on gauging the uncertainty around a parameter; policy experts evaluate actionable alternatives. The media are organized around narratives – actors, plots with beginnings and endings.

Traditional expert judgment, eager to replace the anecdote with the representative, is not favorable to that mode of thinking. Taken absolutely, this refusal is short-sighted. Narrative competence weaves the contextual and temporal dimensions together; it enables media workers to be valuable players in scenario development. Many write with empathy and in style; their expertise comes with presentation skills that can make technical and policy expertise palatable to the intended users.

- **Agency staff:** Many humanitarian agency workers are experts. They were recruited as such, or acquired the expertise, formal or informal, over time, through experience and additional training. Experience grows with deepening exposure to given programs; workers accumulate it also through lateral mobility, which is substantial within and between agencies. The key difference is between insider and outsider perspective. Experts, once ensconced in agency jobs, take on insider perspectives, which demand mixtures of technical and policy knowledge, though in variable proportions. They may lose some of their prior technical fluency or policy evaluator objectivity, replacing them with agency-specific skills and heuristics.

What can we take away from these – largely theoretical – distinctions circumscribing the humanitarian expert role? Humanitarian experts are not exclusively defined by either technical or policy knowledge. Most are valued as experts because of particular personal mixtures of subject-matter, geographic, language, social and cultural intelligence. Their personal traits match the demands of humanitarian agencies in particular situations. Some matches result in long-term positions; others meet short-term needs, serially across places and principals. A minority serve in functional areas that achieve stable syntheses between modern professions and their humanitarian applications, complete with the development and demonstration of autonomous abstract knowledge. Humanitarian logistics exemplifies the successful constitution of a distinct, narrowly technical expertise. By and large, however, humanitarian expertise rests on heterogeneous, syncretic knowledge.

---

10 The concept of “syncretic knowledge” (syncretic = coherent in spite of different sources or opinions) originally refers to knowledge within one given profession that mixes normative and technical styles (Halliday 1985). Malhotra and Morris (2009) exemplify this with the audit profession, whose style is somewhere between those of jurists (normative) and engineers (technical). We apply it to the knowledge produced by humanitarian experts trained in different professions, some with a policy, others with a technical outlook. The concept is worth mentioning here because Malhotra and Morris (and others) have
The consequence is that the talk of “calibrating” such experts is inappropriate. If fifty individuals were identified as “experts in matters of Syria”, neither would they all assemble in a seminar room and submit to a formal test of their knowledge, nor would their collected explanations and predictions be sufficiently comparable in order to be evaluated against actual outcomes. The classic experimental set-up in which a significant number of experts give their opinions on a highly specific question more or less simultaneously may be the exception rather than the rule.

More often, experts contribute sequentially; in fact, there are strong motives for decision-makers to use them sequentially: the information may be complex; the decision-maker or analyst takes time to decode the contributions from individual experts; logistics and calendars force sequencing, which also “spares wasteful consultations whenever uncertainty is resolved early” (Thordal-Le Quement 2016:116). Another motive is the “absence of a commitment ability on the part of the decision-maker regarding future consultation behavior (whether or not he/she will ask for another opinion)” (ibd.) in order to hedge options should the situation change so much that new consultations, perhaps with different experts, are needed.

With so many reservations to consider, we like to keep the definition simple: Humanitarian experts are experts because in the eyes of their principals, the public and other experts they know something that is worth knowing about humanitarian needs, risks and response.

This definition is as inevitable as it is unsatisfactory. It is inevitable because the humanitarian task environment challenges professional boundaries. It is unsatisfactory because it may excuse vicious cycles of low expectations. Notoriously, multi-sectoral gatherings that are supposed to attract agency representatives with a certain level of expertise are confounded with participants whose major qualification is that they are the most expendable on that day.

Between humanitarian and other experts

We conclude the introduction with some general observations on the nature of expert judgment, points so far not covered. We limit ourselves to three points; they are inspired by Meyer and Booker’s “Eliciting and Analyzing Expert Judgment” (2001). These authors, like Cooke (op.cit.), define expert judgment as “data given by an expert in response to a technical problem” (italics ours), but the spirit of their work is wide enough to accommodate policy expertise as well.

Is expert judgment data?
The information that the expert provides becomes data to the extent that it can be separated from the theories and hypotheses that it supports or refutes. This is true of the substantive theory, the theory of the subject matter to which the expert speaks. The data are not independent of the theories that underpin the measurement of the data – an expert predicting refugee flows does not change the definition of refugees laid down in

observed that, as audit firms moved from purely “factual” reporting to risk-based methodologies, they changed their internal organization from strongly hierarchical procedures “towards more lateral team structures and more reciprocal processes of interaction” (ibd., 906). Parallels suggest themselves between “lateral and risk-based” in audit firm and “experts from various backgrounds engaged in scenario-building” in the humanitarian domain.
international law. Meyer and Booker expand the meaning of expert data to include several more components:

- Experts, when given a problem, define its **scope**. In their judgment, they add or exclude issues that non-experts would consider differently, or not at all.
- Experts **refine** the given problem. They break it into specific parts. They identify the key variables that must be measured and connected, the (physical or cognitive) steps that must be taken, in order to solve parts and whole.
- They employ particular **mental processes** to arrive at a solution. One reason why experts differ in their judgment is that a good part of those processes run unconsciously and depend on private meanings.

Decision-makers and analysts may require the experts to make their problem solving processes explicit. The experts may volunteer this information anyway, in order to show their precautions against selection and other biases. Whether this explication should be considered data – as Mayer and Booker do – or just auxiliary information depends on methodological viewpoints. At any rate, it is fair to say that the information that the experts produce takes several shapes: qualitative solutions (text, diagrams), quantitative estimates (probabilities, magnitudes, ratings) as well as auxiliary information on the context of the expertise.

---

**[Sidebar:] Integrated Phase Classification**

In the humanitarian world, the Integrated Phase Classification (IPC) is one of the only well documented methods to generate humanitarian priorities that heavily relies on expert judgment. The IPC is a set of tools and processes to classify the severity of food security situations. It consolidates evidence and perspectives to understand the current and projected severity of the situation, the geographical areas affected by food insecurity, the number of people food insecure, as well as the causes of food insecurity in a given area. The approach of the IPC is to make the best use of what evidence and data are available to ensure that decisions can be made quickly and with sparse information. The process combines a review of secondary data with the interpretation of experts of the available data and gaps. It provides transparency about the confidence levels, and identifies areas for further data collection to improve the quality of the analysis. To know more on the IPC approach, see *IPC 2012 Technical guide*.

---

**When experts disagree**

The patient who receives a severe diagnosis may ask for a second opinion. Privately he hopes that the second doctor will give him a better prognosis, or at least not a worse one. In this regard he may value expert disagreement. Yet, when it comes to deciding on the treatment, the patient prefers agreement. If the doctors agree, the risk of over- or undertreatment appears lower.

In the generic formulation of Meyer and Booker, “many people expect the results of expert elicitation to be reproducible – they believe that experts given the same data will reach the same conclusion”. In fact, they often will not. The authors may go too far when they assert that the “experts may not possess the same data, even if they receive the same briefings and information” because “each expert’s knowledge differs”. The latter part is certainly valid, and is the cause why experts can arrive at different
conclusions from identical premises. The decision-maker, more practically the analyst on his behalf, will therefore want to know how the experts arrived at their respective conclusions.

Expert data packaged such that the paths of creation remain hidden are less trustworthy. To illustrate: During summer 2015, humanitarian planners harnessed three separate streams of expertise to obtain data suitable to estimate the population remaining inside Syria. In the two ground-level streams, sub-district populations came with measures of uncertainty attached to every point estimate. The third stream, based on satellite imagery, supplied point estimates out of a black box, as values of unknown certainty. The absence of confidence intervals complicated the aggregation of the three estimates into one.

While the concept of a population estimate seems straightforward, “expert judgment is frequently sought in situations in which there are no clear standards or well-developed theories”. In such situations, decision-makers and analysts rely on experts for two functions at once: First, they want the clarification of concepts, categories and instruments. Second, they want solutions and estimates expressed in the clarified terms. With one hand, experts work to dispel ambiguity; with the other they seek to reduce uncertainty.

These two endeavors may not progress at the same speed. The social and political environments fixate or dissolve some of the categories that experts, even if they should agree among themselves, cannot control. For example, experts may come up with forecasts of migrants crossing the Mediterranean. They have little influence over how, in the bitter disputes over refugees and economic migrants, the politics of the absorbing countries uses their forecasts (Rellstab 2014). Conversely, the categories of internally displaced person (IDP) and returnee may be undisputed, but in a context of repeated displacement like Syria, it is near-impossible to assign sections of surveyed populations neatly to one or the other category.

So, are expert data better?

Researchers, trying to answer this question, have discovered a paradox. They recognize – as do the decision-makers who use experts – that “expertise is a rare skill that develops only after much instruction, practice, and experience” (Camerer and Johnson 1997:195). Experts “know the state of the art of their field better than nonexperts. For example, novices or lay persons cannot generally describe the state of knowledge in the field; that is, what is not known and what is worth learning” (Meyer and Booker, op.cit., who emphasize) “the expert’s ability to recall greater amounts of visual information. [...] Experts think in more abstract, pattern-oriented ways than nonexperts”. Therefore, in the refining of problems and in solving them, experts can take shortcuts that novices do not know or feel too insecure to take. Experts, when we only look at the cognitive aspects, save time and data – because “they know more and search less” (Camerer, et.al., op.cit.:204)\textsuperscript{11}.

This should give rise to expectations that experts naturally make better predictions than lay persons. Many studies have disproven this; the experts are not consistently better.

\textsuperscript{11} That claim is a bit naïve. It may be true of the personal functioning of the expert. Institutionally, as we know from doctors’ waiting rooms and their habit to order unnecessary tests, that may not be so.
The Camerer and Johnson article is devoted to finding out why knowledgeable experts are poor predictors. Their key concept is “configurational rules”. These are rules that are based on a limited number of cues. The cues are combined in non-linear ways that experts have learned in training or deduced individually from specific, sometimes even rare cases. The non-linear character was first formulated in medical expert systems: “If the reading of test A is above 70, and that of B below 120, and symptom C is present, then the probability that the patient has disease X is greater than 90 percent.”

In the humanitarian field, rules of similar explicit and precise nature are rare. An approximate example can be found in the search rule formulated for pilots looking out for distressed refugee boats between Libya and Italy. “If at night there is a line of sight from coastal point A to the flares of natural gas fields B in the sea, expect boats to progress approx. X km towards the gas fields until they lose sight of them in daylight. From this point, search an area in a circle of approx. Y kilometers defined by their likely fuel balance at day break” (pilot testimony on Swiss TV).

More generally, configurational rules are inscribed into the working knowledge of humanitarian experts who have formed them in dramatic experience in one theater. The likelihood that an expert assigns to the outbreak of waterborne disease after a disaster may depend on her observations of epidemics in previous missions. To mitigate the danger of overlearning from rare and extreme cases, “lessons learned” exercises formulate rules that are extracted from, and tested against, the experience of sufficiently many experts. For similar reasons, the optimism to predict humanitarian disasters through early warning systems (which spawned a cottage industry after Rwanda 1994) has grown more subdued.

Thus, if prediction is not the forte of experts, what are they still good at? Different functions have been promoted as their reason for being. One is to provide solutions; experts know how to avoid major errors and approach major difficulties in their domains. Yet, before they even get to solutions, more basically, experts are “indispensable for measuring variables [...] and discovering new ones” that are relevant to success in their field (Camerer et al., op.cit.: 210). The measurement function may explain the surprisingly stable division of labor between needs assessment experts and response planners. Experts also produce more predictions than laypersons although none may be very accurate. Experts may also be able to better develop relevant scenarios, and do so more rapidly, than laypersons can. In response to “Imagine if A occurs, what will likely follow?” experts will likely offer several scenarios, although their probability estimates may not be superior to those made by laypersons if these were prompted with the same scenarios.

The consequence is that humanitarian experts, like experts in other fields, are good at some things, and no better than lightly trained practitioners in others. In this note, we continue this thread in two ways. First, case studies of scenario-building will show that area experts do not predict one particular development; they work out a small number of likely scenarios, identifying for each one favorable conditions, a broadly defined probability and its likely humanitarian impact. Second, in an appendix, we enumerate the most relevant social and cognitive biases to which both experts and lay people are prey, together with some countermeasures.


2.7. **Summing up**

The gist of this introduction can be condensed in two sentences: Humanitarian expert judgment shares some characteristics with that in other fields. Yet, due to the turbulent environment of much of humanitarian action, the boundaries with other forms of knowledge are more porous, and the expertise is less technical and more mixed with policy knowledge.
3. The process of expert judgment

This chapter describes, in various degrees of detail, the major steps that textbooks on expert judgment customarily enumerate for the process from the perceived need for expertise to its ultimate reception by the decision-makers. Process typologies differ among authors, and ours is eclectic. It must be said upfront that those that we have seen in the literature are molded in the spirit of using multiple technical experts simultaneously even when this was not stated in that way. Understandably so; for this arrangement is attractive for the analyst. Either he aggregates the different opinions statistically, treating them as independent observations. Or he sets up suitable conversations in which the experts themselves work out consensus estimates (so-called “behavioral aggregation”).

Situations with a single expert or with several in sequential manner (see above, page 26) are noted, but do not transpire greatly in the choice of analysis methods that the textbooks highlight. The implied ideal is an appropriate number of experts, each supplying a quantitative estimate, together with a measure of uncertainty, in the response to the same carefully formulated technical question. The aggregated estimates (and their uncertainty) can then be fed into equally quantitatively minded models.

From the humanitarian viewpoint, that tradition leaves a gap open. The multi-expert, simultaneous, quantitatively oriented scenario is not the predominant one in this domain. However, a study like ours cannot do justice to the areas that the dominant literature has treated superficially, such as expert judgments collected in qualitative or mixed-method formats.

Phases of the process

Yet, regardless of the type of expert judgment sought, some of the basic process steps are likely universal. There is a background and preparation phase in which the decision-maker, or someone close to him, keenly feels the need to close a knowledge gap. The need is strong enough to translate into the demand for expertise, and to commit time and resources for the purpose. The questions that the experts will have to investigate are formulated during this phase. The decision-maker often delegates the technical and organizational aspects down the line or sideways to persons thought to have a good understanding of the subject matter and of the market for this particular expertise.
The process of expert judgment

**Figure 3: Ideal-type phases of the expert judgment process**

The recruitment of the experts logically should be the next phase. It is not rare, however, to first select an expert and then let her draft the terms of reference. Frequently, the experts, already recruited or while bidding for the job, will be involved in clarifying and refining the specific questions of interest. This refinement continues into the most critical phase, the “elicitation” of expert opinions. The studied choice of the verb “elicit” over easier synonyms, like “prompt”, “obtain”, or “collect”, etc. hints at its special challenges. In Meyer and Booker’s classic (op.cit.) the section of elicitation procedures stretches over 134 pages; just reading the table of contents is an education.

Recording the opinions that the experts give during elicitation may seem to be a mere, taken for granted, sub-activity. This view underestimates its importance. Depending on the problem and the structure of the information sought, the opinions may arrive in the shape of arrays of statistical results, or of large sets of data that an expert extracted from other sources, together with her estimates. Particularly in a multi-expert setting, this information needs to be recorded in a database appropriate for the purpose, including for the people (analysts, other experts) who will work with it before the results are fed to the decision-maker. The epidemiologists and other specialists who collected samples from Ebola patients must have relied on protocols tailored for rapid genetic analyses by specialized laboratories overseas. Experts in needs assessments that involve community or household surveys may have a variety of templates at their hands. These elements need creative adaptation to local conditions, even in the mundane details of collating and storing the survey data.

Following the elicitation and recording phases, the stage is set for the analysis of the collected expert judgments. However, “analysis phase” is too broad to usefully characterize “the second half” of the process. Certainly, there is an initial evaluation and exploration phase. Much space and attention is subsequently devoted to aggregation and synthesis – of the judgments across experts if there are several involved, or of multiple estimates by one expert if she stops short of combining them herself. The aggregation challenge, in terms of methodological diversity that it has engendered, is on a par with the elicitation knowhow. The two have in common the necessity to carefully characterize the uncertainty, at first of the individual opinions, then of any of their aggregates. Without controlling the uncertainty, inferences from expert opinions are much less valuable, not to say worthless.

The technical literature on expert judgment tends to conclude at this point. This suggests that once the inferences from the expert activity have been cleaned of the defective and highly uncertain ones, the valid ones will take care of themselves. They will reach the decision-maker unaided and straightforward. This is unlikely – except when the decision-maker himself is a sterling expert of the same field and wants to act on the raw inferences without any further packaging. Else, the charge reverts to the analyst, or perhaps passes on to policy experts, who have better sensibilities in framing
The process of expert judgment - Background and preparation

technical communications in decision processes. The literature on communicating research findings may be helpful; it doesn’t seem to be much assimilated in the technical expert judgment books that we have consulted. Regardless, if by the end of the expertise, its findings are still expected to inform the decision (the world may have move on meanwhile), they have to be recast in ways that the decision-maker finds compelling for both their content and their form.

3.1. Background and preparation

The need to obtain expertise may arise at different speeds, clearly from the outset or “darkly through the glass”, in response to a defined knowledge gap or in vague hopes that a renowned expert will contribute something useful to the organization any time she is available. The situations are highly variable as are the ways to meet the need; the expertise sought may be found in a book, in an employee sent to a specialized training course, by hiring an external person deemed to be the expert, or in other types of resources.

Goals, question areas and questions

In the context of expert judgment, our main interest is with the process that leads from a felt need for new knowledge to the actual use of experts. Textbooks such as the already cited one by Meyer and Booker detail ideal-type sequences. Distinct steps translate the initial perceptions into terms of reference and hence the selection of individuals to supply the expert judgments. But it is useful to remind ourselves that organizations produce and consume considerable expert knowledge constantly, silently, and without specific expert ToR. The modern system of professions and the vastly more efficient access to knowledge via the Internet are the structural enablers. In the ideal-type sequence, several steps precede the recruitment of the experts. The need for expertise emerges from unmet objectives and/or deficient processes of the organization. If these are not yet explicit, they have to be clarified. They have to be clarified at least to the point of indicating an area of deficiency or opportunity about which the organization currently lacks sufficient knowledge and skills. This recognized gap supplies the broad goals of the expertise. Obviously, it is not enough to write specific ToR. Meyer and Booker have the organization go through two more critical steps: selecting the general question areas, and identifying the questions that the experts will be asked.

---

12 This literature is fairly large and multi-disciplinary, perhaps with the health sciences as the largest contributors (see, e.g.: Pineault, Lamarche et al. 2010). Lavis et al. (2003:244) argue that “Five questions – What should be transferred to decision-makers? To whom should research knowledge be transferred? By whom? How? With what effect? – provide an organizing framework for a knowledge-transfer strategy”. One may assume that these generic questions are the same for thinking about the translation of expert judgments to decision-makers. These authors assume that researchers and decision-makers are in a position to cultivate long-term relationships: “Over longer periods of time, two-way ‘exchange’ processes that give equal importance to what researchers can learn from decision-makers and what decision-maker can learn from researchers can produce cultural shifts. A decision-relevant culture can be created among decision-makers. .. . Such cultural shifts can facilitate the ongoing use of research knowledge in decision making, not just one-off uses” (op.cit., 227). The potential for such relationships between experts and decision-makers in the humanitarian world may vary more widely. The literature on “communicating uncertainty” is vast; Google Scholar returns over 4,000 references. Here are two recent instances from the humanitarian field: Dalrymple et al. (2016), resp. from research into climate change and disasters: Khan and Kelman (2012).
The process of expert judgment - Background and preparation

The question areas are the themes, topics or issues that the experts will cover. There will likely be several within the stated goals while others of potential relevance will not be noted or will be explicitly excluded. Thus, for example, experts on cash assistance in emergencies may be hired by an agency to advise its program in country X on technical integration with banks and telecoms as well as on cost efficiency. Their ToR may not include beneficiary selection and monitoring, which the program believes to be adequate already.

Questions are detailed and specific – commonsensically, statements that can end in a question mark, inviting a specific answer. Meyer and Booker offer this test: “Something qualifies as a question if the expert finds it sufficiently specific to answer. If an expert cannot answer the question in its present form, it probably resembles a question area more than a question” (op.cit.: 57). Elaborating on the above example, a specific question would be: “For the assumed total program cost of $100,000, how many kg of staple grain would the beneficiaries be able to obtain – A. under traditional in-kind provision, B. using cash assistance?”

Looking only at final expert ToR, we may easily forget that the elaboration of goals, question areas and specific questions is an evolutionary process. It does not usually happen in one short burst of inspiration and instruction, by one decision-maker or his/her agents, and in uniform detail and clarity. The process moves through rounds of internal review and negotiation. It may be delayed by internal conflict (which the prospect of calling external experts may accentuate) and the solicitation of sponsors who will pay for the expertise.

The result is considerable variation in the feasibility of goals, question areas and questions. To revert once more to our example, some in the agency seeking cash assistance experts may be aware that past in-kind assistance used to depress incomes of local farmers. They anticipate that cash assistance will have the opposite effect on grain prices, increasing the cost of living for non-beneficiaries. They would like the experts to model the social efficiency on balance. Others may think that this adds complexity unhelpful for practical decisions, and yet others do not want the possible impact on staple prices discussed for political reasons.

Decisions before experts are hired

In order to better understand and control the variability, Meyer and Booker offer some practical planning guidance, which we reformulate here to some extent:

- “Determine whether the objective is to gather the experts’ answers or to gather their problem-solving processes”. In the first option, the priority is to stop knowledge gaps with specific technical answers, regardless of how the experts arrived at them. In the second, the organization wants to learn the “how-to”, most probably in order to be able to replicate the process with its existing workers (or with cheaper local experts and technicians).
- “Compare the complexity of the areas and the questions”. Complex question areas demand more work hammering out the questions to ask. They pose more numerous questions about the various aspects of the area as well as questions about possible, probable or necessary associations among the aspects.
“Assess the magnitude of the data required”. The experts will base their answers on data (as well as on the models in which they connect the data). Some relevant data already exist; other data have yet to be generated. These may come from primary data collections, secondary analyses or from the retrieval of data to which the experts had access in earlier applications. Regardless, in preparing for the expertise the detail (measures, sample size), quality, confidentiality, cost and time-to-readiness of the required data need critical evaluation. This by itself may be the preliminary subject for an advisory expert, with a keen eye to the timing of the decisions that the final expert judgments need to inform, the sequencing of the operations, and the subsequent ability to aggregate the answers of the various experts if several will be recruited.

Beyond the effort to collect the necessary data, a kind of cost/benefit analysis for the entire exercise is needed. How many external experts are needed in order to produce meaningful detail and variation in their judgments? What kinds of personnel, how many and for how long, are needed to elicit, record and process the expert judgments? What will this cost in money, time and displaced normal activities? Comparatively, if the organization did without the experts, how much would it save, and how much would it lose in terms of competence and opportunity? For example, would the organization risk annoying a donor by not conducting an appraisal using experts from a suggested source?

The evolutionary back and forth among goals, question areas and questions has implications for the involvement in the preparatory phase among the various partners. These include the sponsors (agencies funding the exercise, decision-makers using the expert judgments, groups interested to acquire skills from experts), the affected personnel (managers, analysts, data collectors), and the experts themselves.

Degrees of involvement in the planning

Meyer and Booker elaborate, in considerable detail, various scenarios of the relative involvement of partners. Space reasons prohibit detailed discussion here. In essence, the scenarios are distinguished by the degree to which the external experts assist in the formulation of question areas and questions before their judgments are elicited. They differ on other dimensions as well, notably whether one or several experts are needed, and, if several, whether they need to work in a group or individually. The group format will raise scheduling, cost and logistical problems, but may have advantages of quick turnaround, cognitive consensus, and spontaneous discovery. The authors make a subtle, but important difference regarding the experts’ role in preparations:

- The experts do not participate in the formulation of question areas and questions.
- The experts receive question areas and questions, and are asked to refine the questions.
- The experts receive the question areas, and formulate the questions within them.
- The experts are significantly involved in the formulation of both question areas and questions.
“Refinement” is seen as a significant part of the experts’ work throughout the process. In terms of needs assessments, for example, one would think that experts “refine questions” when they build process models and measurement models, and then translate these back into questions (and hence into questionnaires and databases) that make sense under the given goals and question areas.

Similarly, in the ToR it must be stipulated in what detail the experts will leave their judgments and their data with the organization that hires them. Will they only submit their naked technical answers (such as point estimates and confidence intervals)? Will they dress them up with short rationales? with longer context elaborations? Are they required to submit their problem-solving data and programming code for potential replication?

The multitude and seriousness of the points that the preparation phase must address make one thing clear: Creating the stage for experts to reduce knowledge gaps is no casual matter. It is not your staff retreat, prizing informality and relaxation, with a few external speakers invited to lecture on topics of current interest, without enforced coherence. On the contrary, preparations require common formats, discipline and stability. Whether humanitarian organizations, operating in turbulent environments, can provide the necessary frameworks and infrastructure will later show up in the quality and usability of the experts’ work.

3.2. Recruiting experts

The search light that sponsors, decision-makers and analysts beam out for candidate experts may sweep the near and the far. It may stretch all the way from colleagues / acquaintances / friends “who know something” to individuals far outside current social networks lit up only because of their notability in the relevant field. Commonly, those considered seriously are somewhere in-between – persons that the recruiters or somebody close to them vouch for, and who at the same time have the kinds of professional credentials that matter in impersonal markets for expertise. There is an in-built tension between social control over peers at hand and harnessing depersonalized knowledge from afar.

Reassure or irritate?

The tension is reflected in the choice of experts who reassure vs. those who irritate. An organization may recruit experts expected to reduce uncertainty. It may also seek experts to increase uncertainty strategically. Experts reduce uncertainty when their judgments close a knowledge gap in an otherwise accepted and understood process or model. They increase uncertainty, at least temporarily, when they are invited to reformulate processes and models outside traditional shapes and bounds. The sponsor or decision-maker may call for such expertise deliberately and, not rarely, in defiance of those vested in the tradition. They may suspect that candidates who emphasize their familiarity with the organization have already been coopted to established views and biases. Conversely, groups of subordinates may argue that fresh experts create undue risks and burdens, by ignoring history, culture and tacit routines, and needing more support. They may warn that newcomers may not sufficiently understand decision-making rhythms or likely events that may upset normal rhythms; in this view, fresh experts seem more at risk of missing critical deadlines.
The process of expert judgment - Recruiting experts

The two distinctions – acquaintance vs. stranger; uncertainty-reducing vs. increasing – precede other useful distinctions such as those made in the normative expert judgment literature. On one side, they are still part of the needs identification; on the other, they pre-determine the type of experts that the sponsor or decision-maker looks for in the recruitment.

Meyer and Booker list three requirements for experts to be qualified for the job:

- They must have the required substantive knowledge of the field, i.e. have experience and master the relevant concepts and theories.
- They must have the required formal knowledge (which the authors, for some reason, call “normative expertise”), i.e. must command the techniques, instruments and models needed to collect and process the substantive data.
- They must have those kinds of knowledge in the required detail (“granularity”). One cannot entrust one part of the problem to high-flying macro-theorists while for other parts experts are hired to find answers starting from micro-foundations.

How these requirements can all be met in the recruitment may pose challenges of its own. The different types of knowledge may be amenable to division of labor between experts, such as when a culturally oriented area expert is teamed with a survey specialist.

For other purposes, (e.g., forecasting the dynamic of an epidemic given limited data on its current spread), the participating experts may need to command similar sets of skills and theories in order to interact fruitfully.

We limit this section to one more aspect: the question of the optimal diversity that the recruiters should envisage among the experts that they are going to put to the task.

**Diversity among experts**

Generally, diversity is valued, within the framework of the expert task. Multiple experts produce more meaningful distributions of qualitative opinions or of quantitative estimates than single experts can. Outliers among numerous judgments are easier to recognize and may lead to the anticipation of – dreaded or hoped-for – “black swans” (Taleb 2007). The diversity, however, is productive only under certain conditions. If the experts are to reduce uncertainty, then those recruiting and supervising them must impose common formats in which they can aggregate the judgments (or have pieces of the same puzzle all falling in place). If the recruiters look for experts to strategically increase the uncertainty, they must broadly know what kinds of irritation to the received wisdom the decision-maker can handle. For example, it is one thing to open up for changes in the core processes of a program, and quite another to invite experts that will insist on considering also its outright closure.

Meyer and Booker recommend some optimal numbers of experts. These limits vary by elicitation methods. “If a face-to-face meeting is involved, we recommend having from five to nine experts for each interviewer available to moderate the sessions” (op.cit.: 87). Fewer participants would not produce enough diversity; more would likely struggle with adverse group dynamics (“follow the leader”). For experts that do not meet physically, the ideal range may obey other considerations. If merely numeric estimates are sought, and the marginal cost of adding another expert is low, the upper limit may be much higher, set by timing and representation concerns. However, if every expert is
to supply a detailed rationale of her estimate, the effort to assimilate this much information will force a lower limit on their number.

As noted earlier, perspectives on technical and policy expertise differ; and this may play out in the recruitment of experts. Technical experts focus on narrowly defined questions whereas policy experts keep in mind the diversity of stakeholder views. The former are valued for their codified knowledge, which the recruiters can ascertain through CVs and publications; the primary concerns are not about institutional affiliations. Policy experts are more likely to be selected in an effort to ensure broad institutional representation.

How these various factors and considerations interact in the recruitment of humanitarian experts is difficult to capture in a few observations. Plausibly the division between technical and policy experts is not as sharp as in other domains. For example, designers of sectoral needs assessments may be able to draft questionnaires initially in relative isolation from outside concerns, in a largely technical posture. Subsequently, at review meetings, representatives of other sectors or from competing agencies may flood them with requests to include topics from the borderlands between the sector in point and each of the invitees’ jurisdictions. The resulting final instrument sometimes is a less than ideal compromise between technical coherence and inter-agency buy-in.

We have made most of these remarks with a scenario in mind in which multiple experts are recruited and work concurrently. As we observed earlier, sequential use of experts, and therefore sequential recruitment may be more common. As Meyer and Booker note dryly: “Most selection schemes are based on having experts name other experts. ... The [recruiter] starts with a few known experts, collects additional names from them, and repeats this process until more names are gathered than are likely to be needed” (ibd., 88).

3.3. **Eliciting expert judgment**

**A special type of communication**

All communication among humans consists of an amount of information, the act of transmission and the understanding by the receiver. The challenges of selecting the information, transmitting it effectively and verifying how it is understood have been extensively researched in clinical settings and in survey research (Tourangeau, Rips et al. 2000). The latter has developed prescriptive guidelines on the design of survey questions; underlying them is a four-step model of how the respondents understand and answer questions (Groves, Fowler et al. 2004:202):

- “Comprehension (in which respondents interpret the questions)
- Retrieval (in which they recall [from memory] the information needed to answer them)
- Judgment and estimation (in which they combine and summarize the information they recall or substitute an estimate for a missing element)
- Reporting (in which they formulate their response and put it in the required format).”
The potential for inconsistency and bias in question-answer situations is well known. It is no surprise that the expert judgment literature is replete with prescriptions for the appropriate ways of talking to and hearing from experts. As noted earlier, already the linguistic convention of referring to the stimulation and recording of their judgments as “elicitation” points to a special situation. It signals the demanding needs for preparation, discipline and control.

These are mitigated by several circumstances. Experts know more than randomly selected members of the general public. Many are habituated to intellectual discipline. Thus experts, we assume, submit more readily to formats that would be too artificial for most survey interview modes. They can be motivated to adhere to the “specially designed methods of verbal and written communication” through which expert judgments are gathered (Meyer and Booker, op.cit.:9).

The process of applying such methods is what these authors (and we in this note) understand by “elicitation”. It is opposed to judgments gathered unconsciously, informally or with license for individual experts to contribute in their own chosen formats.

To illustrate the relevance of planned elicitation for humanitarian analysis, we first present a case study about how objects were identified based on a relatively simple distinction for which a hierarchy of competent experts was available.

[Sidebar:] Rapid identification of mine-contaminated communities

Following the so-called Ottawa Treaty banning anti-personnel mines\(^\text{13}\), a substantial number of mine action NGOs collaborated in the “Global Landmine Survey”, coordinated by the Survey Action Center based in Washington DC (Disch and Karlsen 2004). The Survey produced inventories of communities contaminated by mines and unexploded ordnance (UXO) in a number of affected countries. It classified the identified communities by the severity of the socio-economic impacts. Country survey organizations based at the national mine action centers followed a number of protocols written by the Survey Action Center and revised as more country surveys were completed.

Early on, it was recognized that the survey planners in an affected country needed a list, as complete as possible, of communities suspected of contamination before the enumerator teams would be dispatched to the field. In order to update and expand the information available at the national headquarters, supervisors and field editors, shortly after their basic training, would travel to visit provincial and district headquarters. There they would scout for and interview persons familiar in the local histories and patterns of explosive remnants of war and their impacts. Usually, these were military and police officers, hospital superintendents and medical personnel, workers in NGO serving persons with disabilities and in mine clearance agencies. They were the experts.

Survey Action Center Protocol #7, “The Collection of Expert Opinion” (Benini 2001), brought minimal consistency to the approaches chosen by the country surveys. Estimates of the total number of suspected communities were not sufficient for the

\(^{13}\) Officially known as the “Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction, of 18 September 1997”.
The process of expert judgment - Eliciting expert judgment

purpose. Importantly, the planners needed a qualitative grading of communities that the experts believed were definitely mine-affected and those possibly affected.

Figure 4 is a map that illustrates the results of the process in Thailand, a country then with contaminated communities in provinces bordering on four neighboring countries (plus some further inland). For each province, the map displays the number of communities suspected to have landmine or UXO problems together with a pie chart. The red color stands for communities of which the experts believed that they were definitely mine-affected. Yellow represents the portion of communities possibly affected. The map was the outcome of a three-week spree of visits to regional and provincial headquarters. They took place after the training of supervisors and field editors and after the instrument pre-test (in a few villages definitely contaminated) and before the larger pilot test.14

Figure 4: Estimates of mine-affected communities in Thailand, 2000-01, by province

There are obvious regional differences in the level of certainty with which the experts endorsed their estimates. The firmest opinions were offered on the communities along the border with Cambodia (lower right corner). By contrast, experts in provinces bordering on Laos and Burma tended to be uncertain in their opinions.

Source: Benini (2001, op.cit.:1)

The country survey organizations followed a number of principles:

- **Hierarchical search:** It started with the knowledge and data found at the national mine action centers. Survey staff would then visit headquarters of the next lower administrative level, such as the province. If the province-level experts confirmed the existence of a mine problem, the staff would descend to the next lower level, the districts.

- **Multiple experts:** In every administrative unit that the survey staff visited for this purpose, they would interview several experts. The experts represented different groups and professions; the survey staff met with them separately, to avoid group conformity effects.

---

14 While the majority of freshly minted supervisors and editors were conducting visits, others were engaged in the training of the enumerators during the same period. The “Thailand Landmine Impact Survey” took 299 days to complete, going through four phases: Planning, preparation and office establishment (including the expert opinion collection) (102 days), Expansion, refinement, pre- and pilot test (95 days), Data collection proper (85 days), Analysis, reporting and hand-over activities (59), with some overlap of activities. The duration of country surveys varied with the size of the problem and logistical conditions.
Field verification: Subsequently, enumerator teams would visit all suspected communities. They would continuously update the lists of suspected communities as fieldwork went on (however, where the experts at the lowest tier above the communities were good, few new suspects were added by the key informants in the visited communities).

Safeguards against expert error: For control purposes, the survey visited a sample of the non-suspected communities. When a contaminated community was detected among them (a so-called “false negative”), it triggered a local survey and a search for more such communities in its neighborhood. The assumption was that contaminated communities overlooked by the experts tended to cluster (e.g. in corners of a district that were poorer and enjoyed less administrative attention)\textsuperscript{15}.

Several lessons were learned from the experience of the country surveys:

- The variety of experts from whom the survey elicits opinions should not be narrowed down prematurely. This can happen, for example, if the first contacts in the capital suggest that only a certain type of organizations has relevant information, and consequently survey staff do not contact other organizations at province and district levels.
- As one moves down to the provinces and districts, the number of necessary contacts soars. A decision needs to be made early on to “make or buy” (or do both), i.e. to primarily use survey staff to initiate and handle those contacts, or to rely on other organizations such as the government’s field administration or on existing databases (which may be obsolete).
- With judicious scheduling of recruitment, training and instrument development as well as, in an integrated way, of outreach to experts under strict protocols, it is possible to advance both the representation (sample!) and the measurement challenges (questionnaire, GPS, etc.) in parallel.
- One must carefully distinguish between experts and interested stakeholders, who may have their own agendas.

Although in the end the expert opinion process resulted in lists and statistics of mine-affected communities, the approach essentially was qualitative. The effort was organized around the distinction of certainly affected, possibly affected and, by agreement of several experts, non-affected communities. The distinction allowed the country surveys to economize effort by focusing subsequent opinion collections on the possibly affected.

Of course, none of that was perfect. The logic of the hierarchical search implied certain strict inferences, but these had to be reviewed pragmatically. For example, if a possibly contaminated community was found, during the on-site visit, to be effectively contaminated, logically the encompassing district and province, if their status had been “possibly affected”, turned into “definitely affected”. Such an upgrade could be triggered by the mere presence of a few UXO items – somebody kept a stash of hand grenades – in just one community. By the strict survey rules, it necessitated a sampling for false-negative procedure. However, if the triggering event was manifestly isolated.

\textsuperscript{15} This is a stark simplification of a control process that was devised, taught and supervised in great detail, and was the subject of its own separate protocol (Moulton, Ross et al. 2000).
and of little consequence, the procedure would have been highly inefficient. In countries with poor definitions of what constituted a local community and great logistical difficulties – Mozambique was a case in point –, pragmatic accommodations had to be made frequently.

The general lesson that one may take from this application is that the refinement and testing of qualitative typologies, appropriate for the purpose of the expert judgment, needs to precede quantitative estimates regarding particular types. Yet subsequent quantitative findings may throw up questions about the qualitative assumptions and may force the revision of typologies or greater flexibility in their application.

For the prescriptive part, this section relies heavily on the cited work of Meyer and Booker (henceforward in this chapter “M&B”). Occasionally, where M&B are silent or seem obsolete, we draw on other sources. M&B organize their extensive discussion by the main steps to be followed in any type of elicitation process: (pages 11, 122):

- Review possible elicitation methods and their strengths and weaknesses
- Tailor the components to the project objective, available resources and context
- Practice and train relevant personnel
- Conduct elicitation process and document expert judgment

We condense their rich material into selections of pointers that seem important and sufficient for this note. Our rendition is unsatisfactory in one important aspect: It is crucially important to elicit not only the experts’ best estimates, but also the uncertainty that each expert attaches to her estimate. We will demonstrate the use of one such method in the chapter on aggregation and synthesis. The reader interested in an extensive discussion of “uncertain judgments” may want to consult the book by this title (O’Hagan, Buck et al. 2006).

**Elicitation Methods**

**Planning the design:** The elicitation must be planned in detail “because there are so many possible combinations” of situations and methods. (p. 123). In the humanitarian community, this is often neglected and, if noticed at all, later regretted.

For many specific project requirements, proven methods can be adapted. Analysts choose them “in the belief that [proven methods] will enhance the credibility of their work” (p. 99). With the rapid growth of Internet-based research methods, innovations in expert judgment elicitation should be expected although the novel directions are not obvious (see, for an illustration, the sidebar further below).

M&B (pp. 100 sqq) divide the elicitation into five components:

- the elicitation situation,
- mode of communication,
- elicitation techniques,
- response mode, and
- the aggregation of experts’ answers.
The process of expert judgment - Eliciting expert judgment

Each of these components has different sub-components which can be combined into a specific method. (p. 100). We detail them here, but will elaborate further on aggregation (component 5) in one of the following sections.

Table 1: Elicitation components and subcomponents

<table>
<thead>
<tr>
<th>1. ELICITATION SITUATION:</th>
<th>The setting in which the elicitation takes place. The amount of interaction among experts and between the interviewer and the expert determines the situation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>• Best method for obtaining detailed data and data on the problem solving processes experts use to obtain the results. Avoids potential bias from group dynamics; data can be relatively easily processed and analysed.</td>
</tr>
<tr>
<td></td>
<td>• No or limited synergy between experts. Time consuming.</td>
</tr>
<tr>
<td>Interactive group</td>
<td>• Generates more accurate data, particularly for predictions, and more ideas than the other two situations. <em>Appropriate for solving problems that require originality and insight.</em> It creates additional buy-in and legitimacy of results. If experts are to use complex response modes, a group setting is more appropriate as participants can be collectively trained and guided.</td>
</tr>
<tr>
<td></td>
<td>• Potential for group-think bias. Heavy in preparation, administration and logistics. Strong moderator required, particularly if there are more than seven experts in a group.</td>
</tr>
<tr>
<td>Delphi</td>
<td>A setting during which experts do not directly interact with other experts or with the data gatherer (Wikipedia 2016d). The set-up is designed to avoid group bias. However, “several researchers consider the structured interactive group to work better than the Delphi in terms of avoiding the bias” (p. 103)</td>
</tr>
<tr>
<td></td>
<td>• Limited synergy between experts.</td>
</tr>
<tr>
<td></td>
<td>• At the time when M&amp;B published their classic, they considered the method unsuited to gather data on the problem solving process used to reach conclusions. Newer developments (e.g., Seker 2015) give hope that Delphi methods can guide argumentative expert judgments and do so using modern media such as Web-based cooperation.</td>
</tr>
</tbody>
</table>

[Sidebar:] Digital argument Delphi technique

Under the traditional Delphi method, a panel of experts is requested to provide and fine-tune predictions during multiple rounds of consultations. The Delphi method has four main steps. First, experts individually make a prediction on a certain topic. Afterwards the facilitator aggregates all perspectives and shares the results with the contributing experts. These are then requested to update their predictions. Over several rounds, the facilitator tries to reach a consensus prediction (Linstone and Turoff 1975, Linstone and Turoff 2011).

A variety of alternatives to the traditional Delphi method have been developed, the most famous of which is the “Policy Delphi”, which built on opposing viewpoints and debate (Adler and Ziglio 1996). A rare example of its use in the humanitarian sphere
The process of expert judgment - Eliciting expert judgment

is from the Ethiopian famine in 2000 (Cottam, Roe et al. 2004). Not far from the Policy Delphi, the “Argument Delphi method” has been developed to introduce the ‘interchange of arguments between experts’ to the technique (Seker 2015). Seker piloted a digital way of hosting this type of discussion. His article discusses both the concept and an empirical application. While Sekers’ work focusses specifically on the argument Delphi technique, his findings are applicable to a wide range of expert judgement elicitation methods.

In order to understand the opinion of the crowd on future petroleum prices, Seker introduced the following process online:

**Figure 5: The three simultaneous steps of the Argument Delphi method**

An important step within the process is the rating of arguments. Ratings make it possible to process a large number of arguments, catching and eliminating the poor ones. The rating mechanism largely automatizes the role of the facilitator. Once the experts have provided their opinions, they are requested to rate three randomly selected previous entries. The ratings evaluate the quality of the arguments as well as determine the direction of support or opposition.

The computerized version of the argument Delphi method has several advantages compared to its traditional set-up.

- **Variety of sources**: It can elicit and process a large number of contributions from people with a diverse range of backgrounds.
- **Resources**: compared to its face-to-face counterpart, it is faster, cheaper and easier to implement.
- **Supervision**: It comes with an authorization feature version that facilitates the monitoring of contributors. Bad apples can be excluded with a single click.
- **Flexibility**: It is continuous, as opposed to discrete. The aggregation of arguments can take place at any time.

One of the main challenges faced by Seker was the aggregation of the large number of arguments. He proposes several solutions, including a ‘qualitative marginal
The process of expert judgment - Eliciting expert judgment

selection’, which only looks at the arguments proposed to support either extreme side of the opinion spectrum. With a ‘tournament selection’, arguments are rated according to the quality of the point made. The arguments rated lowest, and therefore likely erroneous, are deleted from the end result. This eschews one of the main pitfalls of on-line elicitation – in a time of widespread fake news and state-sponsored propaganda campaigns, it offers a rigorous method to preserve authentic and quality data.

[Table continued:]

<table>
<thead>
<tr>
<th>2. MODE OF COMMUNICATION:</th>
<th>Means by which the expert and interviewer interact.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-to-Face</td>
<td>- Best mode for obtaining detailed information and appropriate to gather data on the problem solving process.</td>
</tr>
<tr>
<td></td>
<td>- Can be time consuming, labour intensive and expensive.</td>
</tr>
<tr>
<td>Telephone</td>
<td>- Can be quick and light in resources.</td>
</tr>
<tr>
<td></td>
<td>- Not appropriate for detailed or long discussions – a phone call should not take longer than 15 minutes (M&amp;B, p. 168).</td>
</tr>
<tr>
<td></td>
<td>- No synergy between experts.</td>
</tr>
<tr>
<td>Computer-aided or Web-based</td>
<td>- A number of new techniques, such as self-administered Web-based surveys and computer-assisted Delphi methods are increasingly being tried out. An overview is missing. Baker et al. (2014) report an experiment in which traditional face-to-face as well as Web-based elicitation were used, but found no clear indication as to which method was preferable. Better guidance is awaited.</td>
</tr>
</tbody>
</table>

3. ELICITATION TECHNIQUES: Techniques to obtain information on the thought process of experts. This is an essential component if the objective is not only to capture responses, but also the process of how to solve the problem.

| Verbal Report | - Instructing the experts to think aloud as they progress through the problem, which is a good way to capture the problem solving process. |
|              | - It is not suited to group situations and only works in a face-to-face setting. Very time consuming. Some research has shown that the process of verbalizing may negatively influence the expert's problem solving process. |

| Verbal Probe  | - Questioning immediately after the expert has reached a solution. “A quick means of obtaining general data on the expert's reasoning in solving the problem.” Can be used on experts individually or in a group. |
|              | - Written responses to the probe are generally inadequate. Meyer and Booker “consider the verbal probe more likely to induce motivational bias than the verbal report because the probe's questioning can
The process of expert judgment - Eliciting expert judgment

| Ethnographic | • Restating the expert’s responses into questions. This approach is appropriate to obtain the greatest amount of detail on the expert's problem solving processes. Relatively without bias and provides a check on misinterpretation bias on the part of the interviewer.  
• Time-consuming. Can distract experts from problem solving and should therefore only be applied once the expert has responded to the question. |

| RESPONSE MODE: form in which the experts are asked to encode their judgments, including probability estimates, ranks, ratings, pairwise comparisons etc. See Table 2 below for more information on the different response modes. |

| AGGREGATION OF EXPERTS ANSWERS: Process designed to obtain a single piece of information from different sources of data. This step is required if multiple experts are consulted, while the objective is to obtain one result. |

| • Behavioral | • During the elicitation process experts discuss and consider until they reach consensus. This approach encourages buy-in from the experts and legitimacy for the results. It protects anonymity because no individual can be linked to the consensus response.  
• Detailed planning is required, including measures to mitigate bias. Risk of group-think discouraging the expression of minority opinions. It can be time consuming if conflicting expert perspectives persist. |

| • Mathematical | • Easier to plan for than behavioural aggregation. “Different mathematical schemes can be applied in succession to the individual's data, whereas with the behavioural aggregation the process can usually only be done once.” (p. 119).  
• Mathematical aggregation masks rationales for particular opinions and the reasons for disagreements. Perversely, “mathematical aggregation can lead to the creation of a single answer that all of the experts would reject.” (p. 119) |
Table 2: Different response modes

<table>
<thead>
<tr>
<th>RESPONSE MODE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
</table>
| ESTIMATE OF A SCALAR QUANTITY        | Need for training of experts: Variable  
Ease of analysis: High  
- Flexible, easy to explain and use  
- Uncertainty: Popular method: eliciting best estimate, minimum and maximum, which then are used as the parameters of a triangular distribution (see example later under aggregation). (O'Hagan and Oakley 2004)  
- Downside: People are typically not good at estimating absolute maxima and minima. They tend to underestimate maxima and overestimate minima.  
- Decomposition of the question has been associated with greater accuracy. |
| PROBABILITY ESTIMATE                 | Need for training of experts: Medium  
Ease of analysis: Medium  
- Easy to explain and use  
- Established elicitation and analysis techniques are widely available  
- Different expressions are feasible. Example (O'Hagan and Oakley 2004:83) :  
  - Probability: ‘There was a 0.08 probability of being a victim of crime in 2002.’  
  - Percentage: ‘There was an 8% chance of being a victim of crime in 2002.’  
  - Relative frequency: ‘1 in 12 people were victims of crime in 2002’, or ‘80 in every 1000 people were victims of crime in 2002.’  
  - Odds: ‘The odds against being a victim of crime in 2002 were 11 to 1.’  
  - Natural frequency: ‘From a population of 56 million people, 4.5 million were victims of crime in 2002.’  
- Can be time consuming and fatiguing.  
- “Research has shown that people have difficulty correctly translating their judgments into quantities such as probabilities. Training is therefore required in their use.” (M&B, p 7)  
- The theoretical background, including the concept of expert calibration, is demanding. For an extensive discussion, consult O’Hagan et al. (2006) |
## CONTINUOUS SCALES

Experts are asked to select a point or range on a scale which is labelled with text or numbers.

**Need for training of experts:** Low  
**Ease of analysis:** High
- Easily converted to numerical, continuous variables for analysis.
- Most people are reliable estimators when using these scales.
- Developing a continuous linear scale to fit a particular application requires time.
- Care must be taken to guard against biased wording of either the labels or of the definitions of these labels.

---

## PAIRWISE COMPARISONS

A process in which experts rate a set of objects, events, or criteria by comparing only two at a time.

**Need for training of experts:** Low  
**Ease of analysis:** Low
- Most people are reliable estimators using pairwise comparisons.
- Time consuming to elicit all possible combinations.
- Only provides relative data relations.
- Current research shows that people make better relative, indirect judgments, such as with pairwise comparisons, than direct estimates.
- Analysis of more complex situations (many objects to compare) may depend on specialized statistical methods such as the Analytic Hierarchy Process (Wikipedia 2016a).

---

## RANKING OR RATING

Assigning numbers or descriptions to the objects, events, or values in question.  
Rankings compare objects. Ratings map objects to a scale.

**Need for training of experts:** Medium  
**Ease of analysis:** Medium
- Easy to use
- Humans have difficulty keeping more than seven items in short-term memory. Therefore, ranking/rating options should not exceed that number (Miller, 1956).
- The use of rankings and ratings has a history of abuse, particularly in humanitarian information management. Guidance is found in Benini and Chataignier (2014).

---

We illustrate the presentation of response modes with an application that is frequently needed in humanitarian expert judgment – the estimation of a proportion. We explain the elicitation requirement in terms of the subsequent aggregation:

---

### [Sidebar:] Eliciting judgments about a proportion

In needs assessments, it is often impossible to rapidly collect detailed information, such as through household surveys. Instead, persons deemed expert in the problem at hand can be polled quickly and inexpensively. For certain questions, it may be helpful asking each expert independently what proportion of the same population she
believes has a certain characteristic or is affected by a certain event. The experts may base their judgments on different amounts of experience, thought of as the number of independent observations that they have made. The challenge then becomes one of combining the estimated proportions, taking into account that some experts base their estimates on only one or very few observations while others have processed numerous distinct ones.

To make this concrete, let us work with a fictitious example. An earthquake has devastated a district with over a hundred village communities. The quake broke underground pipes that conduct water from pure uphill fountains to village tap stands. Village people without functioning water supplies resort to polluted ponds and streams. The question of interest is the proportion of villages in the district with disrupted water supplies.

An assessment team traveling to the district headquarters is able to interview four officials and NGO workers who separately visited a number of villages and enquired about the water situation. These persons were highly familiar with the water situation prior to the earthquake and are considered experts. Based on mixed impressions from conversations and physical inspection, they proffer their individual district-wide estimates. The team notes the number of villages in which they remember spending time on this question.

Some of the experts’ conversations bore on the situations also of neighboring villages that they did not reach, but these details were not captured. Note that this is one of the finer points of this set-up. The experts did not report the number of visited villages with broken water supplies, but offered estimates for the entire district in the form of proportions. We will revert to this anon.

Table 3: Proportions estimated by four experts

<table>
<thead>
<tr>
<th>Expert</th>
<th>Proportion of broken water supplies in the entire district</th>
<th>Villages in which the expert discussed the water situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td>15</td>
</tr>
<tr>
<td>B</td>
<td>0.50</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>0.50</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>0.80</td>
<td>2</td>
</tr>
</tbody>
</table>

Mean proportion:

Unweighted 0.51
Weighted 0.38

The weighted mean of the individual estimates seems a reasonable aggregate measure, but we want an expression of the uncertainty as well – notably to test whether D’s high-end estimate is within the range of the plausible. For this we need a probability distribution of the proportion. O'Hagan et al. (2006:124-132) discuss several methods to derive it from the knowledge of a single expert. The appropriate distribution, as in our motivating example in the Introduction (page 18), is the beta distribution (Wikipedia 2011a). We take elements from O'Hagan et al. in order to propose a simplified method for combining several estimates, given the number of independent observations on which the experts base their personal estimates. We are interested in the mean and in a confidence interval. We want to be able to do all this in an Excel spreadsheet.
The process of expert judgment - Eliciting expert judgment

The beta distribution is characterized by two shape parameters alpha and beta; its expected mean is \( \frac{\alpha}{\alpha + \beta} \). Let \( p_i \) be the proportion estimated by expert \( i \), and \( n_i \) the number of her independent observations. For the single expert case, O’Hogan et al. (op.cit.: 125) derive \( \alpha = n \times p_i \) and \( \beta = n \times (1 - p_i) \). For the multi-expert case, we expand

\[
\alpha = \sum (n_i \times p_i) \quad \text{and} \quad \beta = \sum (n_i \times (1 - p_i)),
\]

where \( \sum \) is the symbol for “sum of”.

Because \( p_i + (1 - p_i) = 1 \), the aggregate expected mean \( \frac{\alpha}{\alpha + \beta} \) is \( \frac{\sum (n_i \times p_i)}{\sum (n_i)} \), which is exactly our weighted mean in table 3 (0.38 or 38 percent).

From here onwards, the calculations are straightforward once we decide the confidence that we want to have in the combined estimate. Let us, arbitrarily, opt for a modest 90 percent.

Table 4: Combined estimate from four expert judgments

<table>
<thead>
<tr>
<th>Expert</th>
<th>( p_i )</th>
<th>( 1 - p_i )</th>
<th>( n_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td>0.75</td>
<td>15</td>
</tr>
<tr>
<td>B</td>
<td>0.50</td>
<td>0.50</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>0.50</td>
<td>0.50</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>0.80</td>
<td>0.20</td>
<td>2</td>
</tr>
</tbody>
</table>

Estimated parameters
- \( \alpha = 9.85 \)
- \( \beta = 16.15 \)

Excel formulas
- \( \alpha = \text{SUMPRODUCT}(\text{R[-5]C:R[-2]C}, \text{R[-5]C}[2]:\text{R[-2]C}[2]) \)
- \( \beta = \text{SUMPRODUCT}(\text{R[-6]C[1]:R[-3]C[1]}, \text{R[-6]C}[2]:\text{R[-3]C}[2]) \)

Population estimates
- \( \text{Mean} = 0.38 \)
- \( \text{LCI90\%} = 0.23 \)
- \( \text{UCI90\%} = 0.54 \)

By combining the four expert judgments, we estimate that 38 percent of the villages at this time are living with broken water supplies. We have 90 percent confidence that the true estimate is between 23 and 54 percent.

Three of the four experts offered judgments that were within this confidence interval. However, the judgment of expert D is not to be dismissed. She likely formed it after visiting two points in a cluster of highly affected villages. Without her estimate, the mean for the three experts A, B and C would be 34 percent, and the 90\%CI would be [20\%, 51\%]. In other words, B and C, basing themselves on 9 villages only, as opposed to A’s 15, would be barely inside the confidence interval.

We leave it to the reader to replicate the calculations for the set-up without expert D. We believe that the task of combining estimated proportions occurs not infrequently when using experts in humanitarian situations, and thus these aggregation formulas may be useful.

In the elicitation perspective, it is important to obtain a measure of the evidence base for each expert. To obtain the parameters of the beta distribution, we need the number of independent observations. In this illustration we counted the number of villages in which each expert stopped to have conversations and see whether the taps had run...
dry. The information, if any was collected, on surrounding villages did not count as additional independent observations because it may have been strongly colored by the impressions formed on site.

What would be different if the experts had “only” reported the number of villages visited and the number of those among them with broken pipes? For one thing, these would be survey data, not expert judgments. The technical side too is revealing. Assume, as in tables 3 and 4, that the four experts visited 26 villages and found ten thus affected (because $0.38 \times 26 = 9.9 \approx 10$). The sampling distribution here follows a binomial model (Wikipedia 2016b). The 90%CI is [23%, 56%]16, which is only slightly wider than the one obtained via the beta distribution, which is [23%, 54%].

The general point is that the elicitation strategy needs to meet the requirements of the eventual aggregation task.

---

**Tailor the methods to the situation**

M&B (page 8) identify eight factors that influence the decision on the selection of the most appropriate elicitation components:

- “The type of information the experts must provide (only the answers or also how they got to those answers, i.e. the problem solving process)
- The form in which the expert's answers are given
- The number of experts available
- The interaction desired among the experts
- Difficulty of preparing the problems
- The amount of time and study the experts will thus, according to Brodley, need to provide judgments
- The time and resources available to the study
- The methodological preferences of the data gatherers (interviewers or knowledge engineers), analysts, project sponsors and experts.”

**Mixing components:** Different components can be combined to create the optimal method for elicitation. For instance, if the resource-intensive interactive group and the face-to-face mode are chosen, these components can be complemented with less expensive ones, such as an individual interview by e-mail. However, not all mixes are appropriate. Thus, if the individual interview has been chosen at an early stage of the process, this mode must be continued throughout, in order to ensure that the level of detail is maintained. (p. 124 and 125)

**Level of structuring:** Preparations are simpler if only a few experts will be interviewed, especially in a face-to-face setting. The more structure the planners impose, the greater the time it takes to plan and conduct the elicitation. (p. 130). “As a general rule, using one of the following components imposes [more] structure: a response mode and dispersion measure, an elicitation technique, the use of behavioral aggregation, or a documentation scheme. Using one of these components means that there is a plan in place and a specific procedure that will be followed”.

---

16 Calculated in STATA, with the command * `cii proportion 26 10, level(90)`.*
Practicing, testing and training

Practicing: Rehearsals should be done of the following parts of the elicitation process (p. 154): the briefings given to the experts, the elicitation procedures, the documentation, aggregation, and entry of the data into a model or database.

Pilot testing: Pilot testing involves a number of experts who are similar to those who will participate in the eventual elicitation process. Pilot testing aims at obtaining expert feedback on material and procedure (p. 154). It focuses on the expert’s understanding of the questions and reassures the planners that the response mode, elicitation process and documentation format will work.

Training: In certain settings, training of the project participants is required before the expert judgements can be acquired. This is particularly important when:

- “Different people plan the elicitation than those who will conduct it.
- The persons who will perform the elicitation and documentation feel uncomfortable with the assignment
- More than one person or team will gather and document the expert data” (p. 157)

The training should cover all those components that are complex and prone to error and confusion, including the elicitation procedures, aggregation of responses and entry/documentation of the data.

While Meyer and Booker recommend conducting rehearsals as well as a pilot test as part of the training, this could result in confusion. Confusion is likely to result if key elements are modified after the pilot test, and the facilitators wind up being briefed with obsolete guidelines. It seems more appropriate that the facilitators should be trained after rehearsals and pilot test in a small group have been done and evaluated.

Conducting the elicitation

After the necessary preparatory work, including the mundane tasks of scheduling and confirming the meetings with the experts, these are the steps in which the elicitations proceed (B&M, p. 168):
The process of expert judgment - Eliciting expert judgment

### Table 5: Conducting the interview and moderating the group - step by step

<table>
<thead>
<tr>
<th>Individual interview situation</th>
<th>Interactive group situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>· Interviewer introduces him or herself</td>
<td>· Distribute relevant materials</td>
</tr>
<tr>
<td>· Start with questions on the expert's professional background</td>
<td>· Introduce the meeting facilitator, project staff and experts</td>
</tr>
<tr>
<td>· Give the expert some sample questions</td>
<td>· Review the project’s purpose, schedule and elicitation procedures</td>
</tr>
<tr>
<td>· Brief the expert on potential biases</td>
<td>· Give experts sample questions to work with</td>
</tr>
<tr>
<td>· Give the expert the set of questions and verbally go over any instructions</td>
<td>· Brief experts on biases</td>
</tr>
<tr>
<td>· Ask if the expert has any questions</td>
<td>· Ask if the experts have any questions</td>
</tr>
<tr>
<td>· Tell the expert that she can begin</td>
<td>· Start the discussion</td>
</tr>
</tbody>
</table>

**Repeat instructions:** During elicitation sessions of more than 30 minutes, people often forget the instructions provided at the start of the session. (p. 137) During longer sessions, participants should therefore be regularly reminded of the set-up and parameters.

**Documentation:** During the elicitation, a record of the process and outcomes should be maintained. The type of documentation depends on the objective and can include the expert’s answers (“answer-only documentation”, p191) and or the problem solving skills (“answer plus problem-solving documentation”, ibd). Ensure that the level of detail in the documentation matches the detail required for analysis and reporting. The selection of methods, and other important decisions made during the planning phase should also be documented.

**Warning signs that the elicitation may be going awry**

M&B found that bias is often ignored or not sufficiently managed (p. 148). There are several signs of bias which should be monitored during the elicitation process. Group bias can for instance be recognised if none of the experts voices a different opinion. Suspect anchoring bias when an expert receives additional information from experts or from other sources during the elicitation but never waivers from her first impression. (p. 149)

**Reluctance or confusion:** “One of its first signs is the experts' reluctance or confusion to provide their judgments. They may mutter amongst themselves or they may openly criticize the elicitation procedures and refuse to give their data.” (p.181) Participants should not be forced to provide their expert opinion. The moderator should politely ask why they are reluctant or confused and then try to address the stated causes.

**Disagreement:** Some view the whole exercise as a failure when experts did not manage to come to a consensus. However, conflicting opinions do not necessarily point to a weakness in the process. According to M&B: “determine if a problem truly exists—in other words, determine if the disagreement resulted from a weakness in the elicitation,
for instance due to ambiguous questions which are interpreted in different ways, or from the natural differences between the experts.” (p. 184 - 185). The disagreement may arise also because the experts work with different data; and from these emerge conflicting judgments. All of this is easier to detect and remedy if the problem solving process has been captured. If not, and depending on the cause of the disagreement and the resources available, additional conversations with experts might be required to resolve unexpected inconsistencies.

How much works for humanitarian experts?
These warnings remind us that the personal competences of the experts and those who work with them are one thing. The ability to harness expert knowledge effectively, so as to answer the questions that are at the center of the elicitation effort, is another. The “psychology of the survey response”, to which we referred at the chapter beginning, operates on the time scale of a pair of “question and answer” within an interview; admittedly, respondents will understand some of the questions in the light of the preceding ones and of the answers that they have already given\(^\text{17}\). Still, the time scale is short. An expert judgment will often be produced as the result of greatly more involved operations, will take more time and require the expert to use books, databases, computer applications and the opinions of yet other experts. The elicitation adds another layer of complexity when it addresses several experts, either separately and individually, or in group meetings.

It is, therefore, appropriate that M&B should develop the elicitation section in such length and detail. Much of their reported research and of their prescriptive advice is pertinent for the work with experts in the humanitarian domain. Yet, the reader cannot shake off a suspicion that there is something congenitally incongruous, as seen from the perspective of humanitarian operations. It is as though the nuclear power plant and its safety experts were radiating from the pages, in a technically complex, yet thoroughly regimented low-probability, high-consequence environment.

The humanitarian world is almost the obverse, with lower technical and higher social contingencies, and an environment that is more turbulent. This turbulence is, almost by definition, powerful in sudden-onset disasters. It may be more placid in deadlocked crises and even more so in preparedness settings. Yet given the high turnover of humanitarian personnel, turbulence is likely to remain significant in most places and times. Ultimately, it makes the work of experts and the ways of eliciting their judgments less pliable to the kinds of detailed, disciplined arrangements that M&B painstakingly laid out for us.

\(^{17}\) For guidance, see, e.g., ACAPS (2016b), available at https://www.acaps.org/sites/acaps/files/resources/files/acaps_technical_brief_questionnaire_design_july_2016_0.pdf.
Aggregation and synthesis
4. Aggregation and synthesis

4.1. Overview

This chapter has two objectives. First, as readers may expect, after information has been collected from multiple experts and properly recorded by the analysts, the next step is to “aggregate” it. Aggregation is an operation normally used in quantitative contexts; an analyst reduces the estimates that the experts have produced of the same variable to one combined value. Ideally, the aggregation produces also a measure of confidence or concordance, such as the confidence interval around the estimate of a continuous variable or a proportion.

“Aggregation” is liable to cause terminological confusion. In the sense that is familiar to most readers, aggregation is the operation of computing statistics of an encompassing category or unit from those of its constituents. Examples include the consumption of fruit as the sum of apples, pears etc. consumed, or the population of a country as the sum of the provincial populations. We will refer to the latter kind as geographic or administrative aggregation; its technicalities are relevant in two of the aggregation methods that we describe.

In other situations, the aggregated expert judgment refers to the same object as the individual judgments; there is no upward inclusion, but rather a horizontal “sort of” averaging operation. At this level, two varieties are distinguished: behavioral and mathematical aggregation. The first employs interaction and sharing of the individual experts’ judgment; we rarely refer to it. The mathematical flavor is performed by the analyst on the basis of separately obtained judgments. Most of the material that follows belongs here. Where analysts use both mathematical and geographic/administrative aggregation, we try to reduce confusion by referring to the former as “combination of expert judgments” and to the latter as “aggregation”.

The literature on aggregation is large; we present an eclectic selection of situations and methods that we expect to be relevant to the work of analysts and experts.

There is an equal need to combine qualitative expert judgments. The appropriate term for this case is “synthesis”, rather than “aggregation”. Synthesis does reduce information, but not in the same straightforward sense as in aggregation. Rather, the information is transformed. In terms of sheer length, the synthesis report may be shorter than the combined expert judgments in text form, and in this special sense the information is reduced. Typically there is no reduction to just one statement, assertion or new hypothesis. We speak of transformation because the work of synthesizing the experts’ contributions goes through several steps that, together, produce more than a mere summary. There is no pertinent synthesis literature from the humanitarian world; we find guidance from outside, from an unexpected quarter.

Second, and this may come as a surprise, we want to create a basic familiarity with Bayesian ways of thinking. Bayesian statistics may be an arcane field for most people, including for the authors of this note; we limit ourselves to a simple exposition of Bayes’ rule and an intuitive grappling with some conceptual notions and qualitative applications. We think this is necessary for two reasons: The Bayesian revolution in statistics is sending waves into fields of inquiry that surround expert judgment
methodologies. Expressions like “updating expectations” and “proper consideration of the base rate” are increasingly common; they originate in Bayesian concepts. Closer to expert judgment itself, a decision-maker who hires experts may consider himself/herself an expert; the estimate that guides a final decision is likely to be a weighted average of his/her own and the experts’. The prior assumptions of the decision-maker thus should be factored in; Bayesian thinking is all about how beliefs change in the light of new evidence such as that produced by the experts.

Recently, Bayesian political scientists have claimed that their methods advance the integrated analysis of quantitative and qualitative data (Humphreys and Jacobs 2015). Evaluation methodologists have since merged their approach with process tracing methods (Befani and Stedman-Bryce 2016). They use this combination of approaches to test how evidence confirms or destroys assumptions made in theories of cause and effect or in questions posed to experts. Those developments advocate that humanitarians talking of expert judgment may want to be at least minimally familiar with this statistical revolution.

We discuss the following methods:

**Quantitative expert judgments**
- Scalars (real-valued unconstrained variables)
  - When the experts state their uncertainty (triangular probability distributions)
  - When experts give only point estimates, but the experts are the same for all objects (Beroggi-Wallace)
- Proportions and probabilities
  - The observation bases of the individual experts are known (beta distribution)
  - The observation bases are not known (Bordley’s formula)

**Qualitative expert judgments**
- Three-step synthesis (ordering, comparing, integrating)

**Bayesian reasoning**
- The basis: Bayes’ theorem
- Updating beliefs on the strength of new evidence
- Process tracing and cause-effect testing

Because the transition to Bayesian reasoning is more natural from the quantitative aggregation section, we place the qualitative synthesis section first.

### 4.2. Synthesis of qualitative expertise

The literature is not very forthcoming on qualitative expert judgments in the first place, and even less so on their aggregation. Frequently “qualitative expert judgment” (or opinion) itself is misleading; authors may use the attribute to connote the opposite of “scientific measurement”. Even when experts give their judgment in genuine qualitative form (textual propositions), “aggregation” is a misnomer in denoting this context. For we are not mapping several judgments to one value on a quantitative scale. Rather, a potential large number of qualitative propositions are reorganized, possibly passed through the filter of relevant theories, and eventually condensed into a smaller
Aggregation and synthesis - Synthesis of qualitative expertise

set of – again mostly qualitative – conclusions. It is more appropriate to think of “synthesis”. Finding appropriate guidance and applications for this expert judgment study is a challenge; the first because of the untamed diversity of qualitative research philosophies, the second because there we did not discover exemplary methodological work demonstrated with humanitarian applications.

For a sophisticated and at the same time lucid treatment in the face of that challenge, we turn to a methodological paper from outside the humanitarian literature. Its substantive inspiration – motherhood in HIV-positive women – is of use here only for illustrative examples, not because of humanitarian connotations. What makes Sandelowski and Barroso’s “Metasynthesis of Qualitative Findings” (2003) so useful is its demonstration that synthesis “ought to be more than a mere summary of … findings”, and how this is done.

Basically, these authors prescribe a three-step process. Some of their language is from esoteric qualitative research schools. Substituting more common terms (and dropping the “meta”), we reformulate the essence thus:

- **Ordering:** The analyst creates a list of all findings from every contribution – in our case from the submission of every expert. The analyst orders the list by combining identical findings and assigning findings to thematic sections.
- **Comparing:** The analyst reorganizes findings along conceptual lines. He compares them with those on other phenomena of interest, including some of those absent from the studies. He assesses the uniqueness of findings or their conformity with known patterns.
- **Integrating:** The analyst examines whether any of the contributions already offers synthetic concepts and their usefulness in interpreting the others. He imports concepts from outside theories and integrates the transformed findings in their light.

**Step 1: Collection, abstraction, ordering of findings; frequencies**

Sandelowski and Barroso want the analyst to make a complete initial listing of the distinct findings in every contribution. These findings then are redistributed to thematic sections. Some of these the analyst had met beforehand, as aspects of the research question; others emerged during the listing process. The analyst marks identical findings and condenses multiple instances into one. To illustrate, their motherhood studies synthesis wound up “with 93 statements representing—in parsimonious form—the almost 800 statements of findings extracted from all 45 reports. These abstracted findings were arranged in 10 sections we named to correspond to the topics addressed in the findings pertaining to motherhood” (op.cit.: 156).

Within every thematic section, the condensed findings were ordered by their frequencies across reports. Two points are important. First, no contribution – and hence no initial findings in them – was rejected “for reasons of quality” (p. 155). This is tantamount to saying, in our sphere of interest, that the experts were not weighted. This is turn is akin to long-standing arguments against statistical weighting in medical

---

18 We may add that since the boundary between “the same” and “very similar” is fuzzy, analysts must abstract from findings with insubstantial differences (such as in wordings only) and treat them as identical.
Aggregation and synthesis - Synthesis of qualitative expertise

prognostics (Dawes 1979, Kahneman 2011:226) and in composite measures in general (Bobko, Roth et al. 2007). Second, a finding extracted from two contributions was counted as one if both instances were based on the same study sample.

The frequencies – in how many reports based on different samples a given finding appeared – are certainly of interest to the analyst. But the real strength is in the abstracting and ordering as the basis for the next synthesis step. 19

Step 2: Reorganizing the findings, comparing them in multiple ways

The analysts re-arrange the abstracted findings under headings that bring conceptual relationships to the fore. Frequencies are no longer of major interest; the reorganization aims at showing “the conceptual range of findings and, in outline form, providing a foundation for the development of conceptual descriptions and models, or working hypotheses” (op.cit., 158). The authors call the reorganized arrangement a “taxonomy”, with its domains and sub-headings. This segment exemplifies.

19 On a less important point, we note that Sandelowski and Barroso’s interpretation of frequencies is problematic. They understand the proportion of reports with occurrences of a finding as its “effect size”. This is confusing. This table displays a small segment of the abstracted findings against thematic sections, together with the claimed effect sizes.

Qualitative synthesis - Ordered abstracted findings (segment)

<table>
<thead>
<tr>
<th>Table 1. Frequency Effect Sizes of 10 Mothering Findings in 45 Qualitative Reports of Studies of HIV + Women</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstracted Findings (N = 93)</strong></td>
</tr>
<tr>
<td>Positive features of motherhood</td>
</tr>
<tr>
<td>1. Children were the main reasons to live, fight, get off drugs, care for oneself, and avoid risky behaviors</td>
</tr>
<tr>
<td>2. Whether their children were in or out of their care or custody, being a mother was central to women’s lives: a source of self-esteem, strength, normality, inspiration, pride, hope, joy, sense of well-being, and sense of self as a whole woman</td>
</tr>
<tr>
<td>3. Children were important sources of physical, practical, emotional, and social support and unconditional love to their mothers, buffering the negative effects of HIV</td>
</tr>
<tr>
<td>Negative features of motherhood</td>
</tr>
<tr>
<td>9. The combination of mothering and HIV was physically demanding</td>
</tr>
<tr>
<td>10. Maternity/life is often the context of HIV diagnosis/death in women</td>
</tr>
</tbody>
</table>

These effects are not effects of the HIV status, because there is no comparison, quantitatively, to findings about non-affected mothers. All these frequencies can yield are frequency differences or ratios within HIV-affected mother studies. E.g., the ratio between items 2 and 3 is = (32/45)/(16/45) = 2.0. This is valuable in itself – as an importance measure - , but it is not an effect size in the classic statistical sense (Wikipedia 2017e).
Aggregation and synthesis - Synthesis of qualitative expertise

Table 6: Example of a taxonomy for re-ordering findings – Segment

The taxonomy provides points of departures for several investigations. Besides cataloguing findings for systematic retrieval and comparison, it helps vigilant analysts to see “what is not there but logically ought to be, [thus] potentially allow[ing] more penetrating syntheses” (p. 161). This is not about missing values in the statistical sense. Rather, it may lead to the discussion of theoretically plausible values that never occurred in the actual distribution of an observed variable. Entirely omitted variables may come to mind; they can be included and filled with estimated or extrapolated values if a basis for this exists in, say, secondary data or expert judgment.

Beyond the curiosity about established as well as novel variables, connecting the findings that fall under a taxonomic heading has other benefits. It engages analysts in the “deliberate search for similarities and differences between a target phenomenon and some other phenomenon—not addressed in the studies reviewed—with an apparent or perceived resemblance to it” (ibd.). What do we already know, or have a good chance of finding out from other sources, about something to which a given finding ought to be compared? For example, when experts offer estimates of child labor in various post-disaster zones, how are these correlated with student attendance rates from recent Dept. Education reports? And, finally, comparing the findings under a certain heading to outside information helps to “minimize the likelihood of inflating the uniqueness of a target phenomenon and to maximize recognition of the relationships between phenomena” (p. 161). Do the studies / experts make the case for something unique? Or, rather, by comparing to what is known about a reasonable “control group” (our term here), are the commonalities stronger than the differences?
Step 3: Extracting the contributors’ own syntheses, importing external concepts

In the last step, the analysts perform two tasks. First, they search the individual contributions for synthetic concepts that a researcher or expert already introduced to summarize a vast swath of her findings. For example, in the reviewed studies, Sandelowski and Barroso found one research team interpreting their material as “eternal motherhood”, another researcher hers as “defensive motherhood”, yet another as “protective motherhood”, etc. This segment illustrates “eternal motherhood” as a set of activities aimed to continue the mother’s presence with her children after her expected physical death.

Table 7: Arranging findings around particular contributors’ conceptual syntheses - Segment

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Conceptual Synthesis</th>
<th>Statements of Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bornes et al., 1997</td>
<td>Eternal motherhood</td>
<td>Videotapes are a means to leave a legacy to children</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mothers choose how they will present themselves</td>
</tr>
<tr>
<td></td>
<td></td>
<td>These videotaped legacies are stories in which they give gendered advice, disclose personal secrets, and express guilt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The concept contextualizing these stories is “eternal mothering”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mothers in this study warned their children about how to avoid mistakes the mothers had made, emphasizing the role gender played in their lives and how it shaped their choices and regrets and, therefore, warnings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mothers warned their noninfected children about AIDS as a deadly disease</td>
</tr>
</tbody>
</table>

Source: op.cit., 164.

Taking note of the synthesis attempts that individual researchers / experts have already made of their own limited material offers tentative perspectives and hypotheses for the overall final synthesis. Moreover, partial syntheses of interest can be reinforced, modified or rejected on the strength of the findings by the respective other contributors, in an endeavor that the authors call “reciprocal translation”.

In addition, the analysts seek to import conceptual elements from outside the reviewed studies, in order to enhance their interpretations of the proposed syntheses. Often qualitative syntheses produce paradoxes and dilemmas. These are legitimate; yet in order to resolve them, the analyst appeals to higher-order abstractions. Thus, for example, motherhood can be both “redeeming and damning” – in realizing positive self-images and in leaving one’s children behind prematurely (HIV-infected mothers) or in negotiating the family/career balance (healthy mothers). Both dilemmas agree with the psychological concept of the “double bind” in conflicting messages (Bateson, Jackson et al. 1956). Dilemmas are common in humanitarian decision-making; they occur also in the interpretation of decisions and of entire policies, such as when the forced return of a migrant is an instance of a refugee policy and at the same time a case for concern under protection policies.

[Sidebar:] Addressing paradoxes with higher-level interpretation

Similarly, in a humanitarian context, paradoxical co-occurrences need explanation with more abstract concepts. To illustrate, the World Food Program (WFP) between 1993 and 1995 expanded delivery points in the highly conflictual northern Bahr-el-Ghazal in the Sudan with more airstrips and drop zones. To outwit looting militias, the active centers often changed from one month to the next. Overall, more delivery points meant
that beneficiaries would travel shorter distances. Despite improved relief accounting, however, the uncertainty about the number of people actually served increased.

In part that had to do with the fact that pilots’ reports on air-borne tonnage and beneficiary counts by ground monitors traveled separate institutional routes. More importantly, more delivery points entailed more people attending food distributions who had traveled from other areas with active relief centers. Thus, the more centers were active, the higher the proportion of beneficiaries that were counted multiple times, albeit to unknown extent. Effectiveness and uncertainty grew in parallel, which was the paradox to explain.

Benini (1997), relating uncertainty to the concept of environmental turbulence (Emery and Trist 1965, Ryu and Johansen 2015), simulated the relationship. Using very few observed parameters, he estimated that the proportion of people not served near their homes or trying to receive from several points would typically grow by a factor of two to four when the population served doubled from 200,000 to 400,000.

What three-step synthesis achieves … and what not

In language less esoteric that Sandelowski and Barroso’s, we have detailed the meanings and key operations in each of the three steps of their method. With segments of tables in their article, we illustrate what each step produces. Humanitarian practitioners may find steps 1 and 2 productive and feasible, but may not see the need for step 3 or consider the effort worthwhile.

This is a valid point. It is obvious that the three-step synthesis is labor-intensive. Already steps 1 and 2 demand considerable effort. The choice to take step 3, or not, will depend on additional, largely non-academic considerations. Chiefly, the analyst must decide whether a synthesis effort limited to steps 1 and 2 gives the decision-maker the expected answer. Step 3 is inevitable if the decision-maker wants to have the synthesis recast in a different perspective. He/she may be interested in the synthetic attempts of one or all of the experts, perhaps enriched with some imported concepts. He/she may insist on comparisons with relevant outside “control groups” or on acknowledgments of current big-name thinkers. He/she may want the synthesis re-written in a different philosophical language from the one used by the experts. For an analogy, the efforts required to recast development and humanitarian programs from a needs-based to a rights-based approach come to mind (Stevens 2016 as one of a large literature on this topic). The switch involved a far-reaching philosophical reorientation; for development agencies choosing this path, the effort was major.

When the synthesis goes through all the three steps, “the result is an interpretation at least three times removed from the lives represented in them: it is the synthesist’s interpretation of researchers’ interpretations of research participants’ interpretations of their lives” (Sandelowski and Barroso, op.cit.: 167). The participants’ interpretations constitute the first distancing from their naïve lives; the various researchers do it a second time; the synthesis caps it with the farthest removed interpretation. The authors believe that the process produces an “amplification of data” (p. 154). From an information-theoretic viewpoint, that is not so. Synthesis, like aggregation, reduces information. However, by adding interpretation at the same time, it does expand knowledge.
Sandelowski and Barroso’s article serves us well, as a ready axe to cut a path through the thicket of proliferating qualitative methods. Is it the best tool in the expert judgment context? Impossible to know, short of a much farther flung methodological network! Readers wanting to dive deeper into research synthesis will be well served by the handbooks of Higgins and Green (2011) and Gough et al. (2012), well beyond the brief of our study. Further to note: This is not qualitative data analysis, with its emphasis on coding small text units and building a structure of categories on top (e.g., from a wide range of textbooks: Dey 2003, Bazeley 2013). The primary unit in Sandelowski and Barroso is the research finding, in other words: a proposition that can be true or false.

[Sidebar:] Research synthesis on the impacts of cash transfers

Since the turn of the century, cash transfers – the payment of cash, instead of in-kind assistance or loans, to needy persons – have been rapidly expanded as part of development programs as well as of humanitarian assistance (RCRC 2016). The London-based Overseas Development Institute (ODI) commissioned a review of the large body of studies into its effects (Bastagli, Hagen-Zanker et al. 2016).

At first sight, the 300-page report does not appear to fit the three-step process described in the previous pages. Bastagli et al.’s perspective is rigorously quantitative. In a Herculean effort, the team retrieved more than 38,000 studies and reduced them to a list of merely 201 included in the research synthesis. The criteria for inclusion were chiefly the demonstration of quantitative effect sizes satisfying approved statistical methods.

At second glance, however, it becomes obvious that the study team prevailed in the kind of disciplined triathlon of ordering, comparing and integrating that Sandelowski and Barroso prescribed. The ultimate theoretical framework rests on the distinction of first, second and third-order outcomes. Examples of these at the micro level are listed in this figure; noteworthy is the insight that some outcomes have sectoral correspondences across all three orders (e.g., education, health) whereas others do not easily cross orders under the same traditional sectoral labels (e.g., general household expenditure [1st order] – self-acceptance, etc. [2nd] – Psychosocial wellbeing and social capital [3rd]). This partial-order limitation is a well-known challenge in causal and evaluative studies, also in the humanitarian world.
Aggregation and synthesis - Aggregation of quantitative expert judgments

In the expert judgment context, this study is remarkable also for its use of expert recommendations and snowballing in the search for unpublished studies and the use of text mining for the evaluation titles and abstracts of digitally retrieved studies. Text mining results led to the exclusion of over 18,000 studies. Exclusions were made also for a number of other reasons (e.g., language). Only 639 studies made it to the full-text screening for the evaluation of effect sizes, and another 449 for the studies of other research questions. Only 201 made it into the final research synthesis.

4.3. Aggregation of quantitative expert judgments

“Aggregation” is the operation that turns the quantitative judgments of several experts into one quantity. As noted, at this level, aggregation is of two kinds – mathematical and behavioral. The mathematical variety is performed by an analyst, who takes every participating judgment as an input to a mathematical function. A judgment in this sense consists chiefly of an estimate as well as the supporting rationale, data and algorithm; aggregation primarily means the combination of the individual estimates. Behavioral aggregation brings the experts together to work out a consensus judgment, either by interacting directly or through a supervised arrangement such as the Delphi method.

This section addresses a small variety of mathematical aggregation situations. We discuss two situations in which experts provide estimates of a scalar, a single-standing real number. Population estimates are likely to be the most relevant application in the humanitarian field. Similarly, humanitarian analysts may need methods to aggregate proportions estimated by multiple experts. The need arises when experts provide only
proportions, i.e. when there is no information on the population numerators and denominators. Again, we discuss two relevant situations.

Multiple-expert estimates of a scalar

Several situations may occur. Two important distinctions concern whether 1. the experts provide a measure of their uncertainty or only their best estimates, 2. the experts are the same for all objects judged (except, perhaps, for missing judgments), or they differ from object to object.

The case where experts only provide their best estimates as well as differ from object to object is uninspiring. Think of key informants who provide their individual estimates of the populations in their respective districts, but nothing more. The most that can be done safely with these estimates is taking the district means or medians and then administratively aggregate them by simple addition. We shall not discuss this further.\(^{20}\)

Two other cases are of greater interest:

- The experts state their uncertainty. For each object judged (e.g., a given district), the experts may be different (e.g., the key informants in district A are not the same as in B).
- The experts only give their best estimates. The same experts judge all objects (e.g., overall there are three experts; the same individuals estimate the populations of every district of interest).

Experts stating uncertainty; different experts

O’Hagan et al. (op.cit., 181 sqq.) discuss, under the heading of “opinion pooling”, a principal dilemma that confronts the decision-maker using multiple experts for this type of estimates. We borrow their diagram for a non-technical discussion. The idealized case deals with just two experts estimating the quantity of interest in just one object. The x-axis represents the quantity of interest, the y-axis the probability density.

---

\(^{20}\) We should, however, mention that Douglas Hubbard, in his popular “How to measure anything” (2014:29-30), makes a case for what he calls “The Rule of Five”: With any five estimates of the same continuous variable, we can be 93.75 percent confident that the true median of the entire population lies between the lowest and the highest estimates, e.g. if five experts suggest 30, 60, 45, 80 and 60, the interval is [30, 80].
Figure 7: Two ways for the decision-maker to aggregate the uncertainty of experts

Source: O’Hogan et al., op.cit., 183. Text boxes and arrows by us.

Expert no. 1’s best estimate is 2, no. 2’s is 4. Both express their uncertainty in the shape of normal probability distributions (dotted lines). Expert no. 1’s range of practically relevant values runs from -1 to 5; no. 2’s from 1 to 7. These ranges overlap on [1, 5].

What shall the decision-maker (or analyst) make of these two uncertain estimates? The authors argue that it depends on the beliefs in the ability of the experts:

- The decision-maker believes that both experts are good judges of this quantity “on the basis of the information available to them, but they have different, independent sources of information”. In this case he/she “concentrates on [the values] that are well supported by both experts”. This is the narrow distribution with a best estimate of 3 and a practical range [1, 5] (dashed line).
- Conversely, when the decision-maker doubts the ability of the experts, his/her “beliefs encompass the full range of values that either expert considers reasonable”. The wider distribution prevails. It too has a best estimate of 3, but a wider practical range of [-1, 7] (solid line) (ibid., 183-84).

Note that in the first case the resulting distribution is narrower than either of the individual distributions. This implies that the belief of the decision-maker is stronger than the experts’ individual beliefs. The opposite holds for case #2.

It is worth reformulating these results because decision-makers and analysts sometimes unconsciously follow one or the other underlying logic. Case no. 1 implies that an expert can veto another expert’s range of values, and vice versa. If the decision-maker is cautious to the extent of distrusting the experts’ abilities (case no. 2), an expert can force the expansion of the other expert’s range of values, and vice versa.

There are two practical consequences:
With two experts, their practically relevant value ranges may not overlap. The aggregated best estimate may then be outside either range. Neither expert would consider it feasible. With more experts, this situation is even more likely.

With more than two experts, neither the intersection (case no.1) nor the union (no. 2) of all the practical value ranges are satisfactory for the decision-maker. Most likely, they are either too narrow (or even empty) or too wide. In both cases, the decision-maker should set a confidence level, and the analyst then calculates the confidence interval on the aggregate distribution. Weighting the experts may also be necessary.

The key point is that the form and result of the aggregation do not solely depend on the expert judgments. The decision-maker’s and analyst’s beliefs direct them.

**Practical implementation**

We now turn to a practical implementation that is easy at the elicitation stage and amenable to mathematical aggregation in a spreadsheet program. It works with the kinds of key informants who supply the local expertise in humanitarian assessments. Analysts can manage the aggregation as well as derive reasonable confidence intervals in spreadsheets.

The method is based on the triangular probability distribution (Wikipedia 2015). The distribution is popular particularly among risk management applications. It has three parameters that the expert (or key informant) supplies in response to:

- “What is your best estimate?”
- “What is the minimum below which you will not go?”
- “What is the maximum above which you will not go?”

To illustrate, let us imagine a key informant striving to estimate the population of her neighborhood in a war-torn city.
Aggregation and synthesis - Aggregation of quantitative expert judgments

Figure 8: Minimum, best estimate, and maximum represented by a triangular distribution

Source: Henrion et al. (2012). x-axis title by us.

This person knows something about former census results and about the history of the city since the conflict began. Because of repeated displacements in and out, she is very uncertain about the range of plausible values. Her best guess is that the current population is 180,000. She does not believe that it is anywhere above 230,000, or any lower than 55,000. This minimum is so low because over the years many people have moved out, but she has no firm idea as to how many. If we use a triangle to represent her uncertain beliefs, we can also say that she gives it only a 10-percent chance that there are fewer than 100,000 people in this neighborhood, and only a 10-percent change that there might be more than 200,000. Her preferred number is 180,000 – and thus the distribution is not symmetrical.

Other key informants in the same neighborhood, interviewed by other agencies, may have different beliefs, represented by different triangles. Every triangle has a minimum, mode (most likely value, most plausible value, “best estimate”) and maximum – because the enumerators asked every key informant those three questions.

The mathematical aggregation of triangular distributions will in most cases rely on simulation methods. The analyst can do this in a spreadsheet application, with the help of a user-defined probability function. The sidebar outlines the technicalities for the interested reader. The graph here visualizes an example from an OCHA-coordinated estimation of the population inside Syria in summer 2015. From Taldu, one of 217 sub-districts with population estimates elicited from multiple key informants, three estimates were obtained.

Informant A estimated a minimum of 49,000 residents, a most likely figure of 49,900 and a maximum of 53,000. B’s figures were 53,000, 54,000 and 56,000, respectively. For C, 55,000, 55,136, and 60,000 were noted. Precise figures like “55,136” may have come from earlier counts or combined lists by relief committees.
Aggregation and synthesis - Aggregation of quantitative expert judgments

Figure 9: Aggregation of key informant estimates

Note that only two of the three triangles significantly overlap; A’s and B’s meet in one point only.

Exactly 1,000 random variates were drawn from these distributions, 334 from A’s, and 333 from B’s and C’s each. The resulting aggregate distribution, colored light brown and smoothed for aesthetics, has two peaks because of the low overlap of the triangles. It extends slightly beyond the combined triangle bases because of an adjustment commonly made: The minima and maxima designated by experts are psychologically conservative; therefore, they are stretched outwards by a certain amount.

The mean of the 1,000 random values is the aggregate best estimate. In this particular simulation run, it was 53,940. It lies outside two of the key informants’ ranges (it is larger than A’s maximum and smaller than C’s minimum).

Critical readers may question the trouble of going through the simulation. The simple arithmetic mean of the individual best estimates is a close-by \((49,900 + 54,000 + 55,136) / 3 = 53,012\). Certainly, this value must be inside any reasonable range. Indeed, the 95-percent confidence interval, calculated from percentiles of the 1,000 draws, is a wide [49,412, 58,672]. Wouldn’t it be good enough (and painless) to work with the arithmetic mean of best estimates (53,012), with the interval bounded by the minimum of the minima (49,000) and the maximum of the maxima (60,000)?

This may be so when experts judge only one object (such as Taldu sub-district). However, the more triangles there are, and the more they overlap in the center, the farther will the bounds of the confidence interval be pulled inward from the MINMIN and MAXMAX points. In other words, simulation with confidence intervals produces superior estimates. In addition, while simultaneously working on numerous expert judgments (a total of 1,323 key informants participated from the 270 sub-districts) and
laying the groundwork for the subsequent administrative aggregation, the semi-automated large spreadsheet calculations are immensely more efficient. For the interested reader, further technicalities are outlined in the sidebar below.

In sum, the triangular probability distribution is an attractive model to capture the uncertainty around a scalar quantity of interest. It makes for easy interviewing of experts and key informants, which is an advantage particularly in humanitarian assessments. The mathematical as well as administrative aggregations are flexible in the sense that the number of experts per objects and the number of objects within higher-level objects may vary. The calculations for both kinds of aggregation are manageable for mid-level spreadsheet users, provided they receive templates with instructions.

[Sidebar:] Technicalities of aggregating triangular distributions

As noted, here we are concerned with two forms of aggregation, mathematical and administrative. Behavioral aggregation, a social consensus process, is relevant in other contexts.

A. Mathematical aggregation

- **Prior adjustment:** Every expert’s stated minimum, mode (most likely value, “best estimate”) and maximum define a triangular probability distribution. An adjustment can be made for the fact that in most experts, psychologically, the verbalized minimum is not their absolute minimum, but rather a value close to the 5th percentile of their subjective probability beliefs. Similarly for the maximum and the 95th percentile. The adjustment modifies the extremes such that a set percentage of the probability mass is moved outward from the median. The purpose is to obtain a more conservative (=wider) confidence interval. In many practical situations, the difference with/without adjustment will not be vital. The calculations in Excel are not straightforward. The analyst may in good conscience work with an approximation. If the minimum is > 0, he may multiply it by 0.9 and the maximum by 1.1 and then work with these modified values. The resulting triangle has a wider base and a lower tip (the tip is lower because the triangle area must remain = 1). See Greenberg (2012) for more involved adjustment steps.

- **Random draws:** An analytical solution (a “formula”) of the mathematical aggregation is unwieldy or impossible as more experts participate. Simulation, by means of appropriately large numbers of random draws from each triangular distribution, produces approximate results. These should be “good enough” by any reasonable standard. The simulation work in a spreadsheet is technically simple if occasionally clumsy (pasting formulas over, say, 1,000 columns to the right). Random numbers from a uniform distribution (the function =RAND() in Excel) are fed into the inverse cumulative distribution function of the triangular. In Excel, this user-defined function takes the random number as well as the expert’s minimum, mode and maximum and returns a random draws. It can be copied into a VBA module of the macro-enabled workbook and then is always available:
The handling of this and related functions is demonstrated in the appendix starting on page 173.

- **Draws per expert:** As noted, the experts are different from object to object of interest (e.g., from district to district). If the quantity of interest (e.g., the population) is estimated for several objects, it is good practice to create the same number \( N \) of random variates for each. \( N = 1,000 \) is abundantly sufficient for most purposes. For a given object \( i \) in this set, \( n_i \) experts provide their individual estimates of minimum, mode and maximum. They define \( n_i \) triangular distributions. From each of these, the analyst draws \( N/n_i \) values (with minimal rounding up or down so that they sum to \( N \)).

- **Table arrangement:** Objects and experts are arranged vertically, with the parameters shown in the next three columns to the right. The columns for the triangular random variates follow next. In Excel, it is best to hold the random numbers from \( \text{=RAND()} \) in a different sheet, generated and frozen at the corresponding cells directly behind those of the main sheet. Among the experts of object \( i \), the \( N/n_i \) draws from each of their triangles are arranged in a **sieve**. The sieve ensures that in later aggregations over several objects every object always participates with exactly one random draw in every column. This segment shows two sub-districts from the Syria exercise, one with estimates by three key informants, the other with two. Each sub-district has exactly one random draw per column.
### Aggregation and synthesis

**Aggregation of quantitative expert judgments**

#### Figure 10: Random draws, arranged in a sieve

<table>
<thead>
<tr>
<th>Object ID</th>
<th>Expert ID</th>
<th>Experts / object</th>
<th>Minimum</th>
<th>Mode</th>
<th>Maximum</th>
<th>Random draw 1</th>
<th>Random draw 2</th>
<th>Random draw 3</th>
<th>Random draw 4</th>
<th>Random draw 5</th>
<th>Random draw 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>49,000</td>
<td>49,900</td>
<td>53,000</td>
<td>51,631</td>
<td>50,419</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>55,000</td>
<td>55,136</td>
<td>60,000</td>
<td>57,166</td>
<td>55,918</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>53,000</td>
<td>54,000</td>
<td>56,000</td>
<td>54,003</td>
<td>55,012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
<td>14,000</td>
<td>16,000</td>
<td>18,000</td>
<td>14,853</td>
<td>15,896</td>
<td>15,637</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>2</td>
<td>13,500</td>
<td>15,000</td>
<td>17,000</td>
<td>14,497</td>
<td>15,337</td>
<td>15,163</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sieve can be created by taking the sum of the column and row numbers, dividing it by the number of experts for the object, and showing the value of the inverse cumulative if the modulus = 0, else hiding the value. The Excel formula for this would be:

\[
=I F(M O D(R O W(R C)+C O L U M N(R C), R C 3)=0, I n v e r s e T r i a n g u l a r([ w i t h 4 a r g u m e n t s ) ), ""
\]

In the Syria exercise, with 217 objects and a total of 1,323 key informants and 1,000 random variate columns, the formula was copied to 1,323 million cells, displaying 217,000 random draws from triangular distributions.

For some individual object \( i \) with \( n_i \) experts, all the \( N \) draws can be collected in a summary row below its \( n_i \) rows. This is visually effective for a single object or for very few of them. It is infeasible for any sizeable number of objects. A separate summary table with one row per object and one column per draw is the way to go (in Excel, Pivot tables with 1,000+ fields will not work. Named ranges and the functions \text{SUMIF} \ and \text{INDIRECT} provide a work-around.). This is a screenshot of such a table for the two example sub-districts, with 6 out of the 1,000 random draws shown:

#### Figure 11: Summary table with one row for every assessed object, segment

<table>
<thead>
<tr>
<th>Object ID</th>
<th>Sum of Random draw 1</th>
<th>Sum of Random draw 2</th>
<th>Sum of Random draw 3</th>
<th>Sum of Random draw 4</th>
<th>Sum of Random draw 5</th>
<th>Sum of Random draw 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taldue</td>
<td>51,631</td>
<td>57,166</td>
<td>54,003</td>
<td>50,419</td>
<td>55,918</td>
<td>55,012</td>
</tr>
<tr>
<td>XYZ</td>
<td>14,853</td>
<td>14,497</td>
<td>15,896</td>
<td>15,337</td>
<td>15,637</td>
<td>15,163</td>
</tr>
</tbody>
</table>

- **Statistics:** The mean, median, etc. for each individual object can then conveniently be calculated to the right of the \( N \) random draw columns. This is the last necessary step in the mathematical aggregation. The VBA module in the template should also give code for the probability density and the cumulative probability. These are needed for optional visualizations, of the experts’ triangles as well as of the aggregate distribution. In Excel, which lacks a univariate density smoother, they can be shown as histograms.

### B. Administrative aggregation

#### 1. Sequence of operations:

The analyst will, in many situations, perform a second aggregation, which we call the administrative one. For example, the estimated population of a province is the sum of the estimates for the districts it contains. This is not true of rank-based statistics such as the bounds of the confidence intervals. They are not additive. Confidence intervals are desired
Aggregation and synthesis - Aggregation of quantitative expert judgments

also at the higher levels of province and nation; they are obtained in the two-step procedure below.

In the administrative aggregation, only the calculation of the mean is commutative. That means, for example, it does not matter whether first the best estimate for every district in the province of interest is calculated as the row mean of all random draws for a given district, and after that the provincial estimate results from the column sum of the district estimates. Or, alternatively, first we add all the district estimates (in every random value column), and then obtain the best provincial estimate as the row mean of these. The result is the same as long as we use the same number of random draws for every expert. **Emphatically, the result is not the same for the rank statistics** – the median, minimum, maximum and percentiles needed for confidence intervals. **We first aggregate (i.e., sum) in the columns. In the second step, with the random sums at the higher-level unit, we calculate these statistics row-wise.** Calculate higher-level statistics down the columns from the lower level statistics would be incorrect.

Figure 12 shows the correct sequence. In the example, District ABC (the higher administrative level) consists of the two sub-districts Taldu and XYZ only.

Figure 12: Administrative aggregation and higher-level statistics

<table>
<thead>
<tr>
<th>Lowest-level random draws:</th>
<th>Lowest level statistics, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object ID</td>
<td>Random draw 1</td>
</tr>
<tr>
<td>Taldu</td>
<td>51,631</td>
</tr>
<tr>
<td>XYZ</td>
<td>14,853</td>
</tr>
</tbody>
</table>

**Step 1: Sum by column**

<table>
<thead>
<tr>
<th>Higher level sums, e.g. district population total:</th>
<th>Higher level statistics, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supra-object ID</td>
<td>Random sum 1</td>
</tr>
<tr>
<td>District ABC</td>
<td>66,484</td>
</tr>
</tbody>
</table>

**Step 2: Compute statistics by row**

The commutativity of the calculation of the mean is immediately obvious; for 69,255 = 54,025 + 15,231 (-1 for the rounding error). However, the minimum of the district population draws (65,756) is not equal to the sum of the sub-district minima (50,419 + 14,497 = 64,916). These differences tend to become larger the more objects a supra-object covers. The point matters in the context of obtaining valid confidence intervals.

2. Aggregation to further levels: The same logic applies to any further administrative aggregations, such as provincial populations as sums of district populations, etc., and to statistics at such higher levels.

When the experts are the same for all objects judged

Sometimes the experts do not quantify the uncertainty of their estimates of a scalar quantity. This diminishes the value of their judgments because the uncertainty is harder to contain. However, if the same experts judge all the objects under consideration, the pattern of their estimates can be exploited in order to derive individual weights. In other
words, the higher the weight on an estimate, the smaller supposedly the expert’s error. Obviously, weights are of interest for the combination of estimates only. For each object, the analyst calculates a best estimate as the weighted mean of the individual estimates. In the second step, the administrative aggregation produces supra-object best estimates as the sum of the member objects’ best estimates, again without an uncertainty measure.

One such method has been proposed by Beroggi and Wallace (2000). We present a simplified version. The calculations can be handled in a spreadsheet, which makes it a feasible option in humanitarian information management settings. The sidebar below gives more technical specifics and a small numeric demonstration.

The authors’ basic idea is that, not one, but two data-driven factors should determine the weight on an estimate of object \(i\) by expert \(j\):

- **Agreement**: The relative deviation of the estimate from the mean of all expert estimates of object \(i\)
- **Consistency**: How consistently expert \(j\) tends to offer the highest, or lowest, estimate across the judged objects

One technical feature needs to be mentioned: The agreement score is a continuous measure; the consistency score is based on ranks. Together they determine the weights of the estimates. Experts whose opinion tends to be close to the group mean for most objects and whose opinion ranks vary the least get the highest weights and thus have the strongest impact on the aggregated values.

Thus the method rewards conformist judgments and penalizes outliers. Depending on how the agreement score is calculated for the farthest outlier value of a given object – assigning it zero or a positive value –, the influence of outliers is simply deleted or is preserved in combination with their experts’ consistency scores.

Simply assigning all strongest outliers zero weights seems questionable. If, for example, expert X always offers estimates that are by far the highest, she will be maximally consistent, but in strong disagreement with the rest of her peers. Her consistent outlier position may express systematic bias. If so, the method correctly filters out her influence. But perhaps X knows something than the others don’t. At the least, the analyst should either investigate the causes of, and rationale for, X’s bias or choose a formula for the agreement score that preserves small, but not negligible weights for outlier estimates.

In what situations is the simplified Beroggi-Wallace method appropriate?

- There are at least three experts; a-priori they appear equally reliable; there is not enough information in order to weight them.
- The experts judge independently.
- The experts do not attach uncertainty measures to their estimates.
- Analysts cannot triangulate individual estimates interactively with the experts.
- There are few missing estimates, or, ideally, none. In other words, every one of the experts provides estimates on most or all objects.
Aggregation and synthesis - Aggregation of quantitative expert judgments

To illustrate, in the afore-noted estimation of sub-district populations in Syria, three
major expert sources had to be combined. Two of them provided uncertainty measures;
best estimates were calculated using the triangular distribution method discussed above.
The third, using satellite imagery, worked with an undisclosed algorithm, supplying
only best estimates. For 208 of the 270 sub-districts, estimates were obtained from all
tree sources; for the remaining 62, two sources in varying combinations were available.
The missing rate thus was 62 / (3 * 270) \approx 8\% percent, high enough to necessitate
imputations. Each source was treated as one collective expert. The estimates by the
three “experts” then were weighted in each of the sub-districts, combined and summed
up to a national estimate. Eventually, a confidence level was wrapped around the
national estimate proportionate to its width around the corresponding estimate in one
of the first two methods.

When there are only two experts

Finally, we briefly elaborate on the situation when only two experts judge the objects.
Methods for gauging the degree of their concordance are well developed, going back
to Bland and Altman’s classic “Statistical methods for assessing agreement between
two methods of clinical measurement” (1986), which has been cited more than 36,000
times. Pragmatically, what analysts working in spreadsheets can do is to plot the
difference between the two measurements against their arithmetic means, known as the
limits-of-agreement (LOA) procedure. The LOA graph for the satellite- and key
informant-based population estimates for 211 sub-districts in Syria exemplifies this
graphic approach.

Figure 13: Limit-of-agreement graph for population estimates by two experts

The limits of agreement are far apart; the concordance is low. The mean difference in
the population magnitudes is -0.108, i.e. on average the satellite-based estimates are
1 \times 10^{-0.108} \approx 22\% percent smaller than the key informant-based ones. The standard
deviation of the magnitude differences is 0.496. In Excel, the 95\% limits of agreement
in this case are calculated as =NORM.INV(0.025, -0.108, 0.496) = -1.08 (lower limit)
and =NORM.INV(0.975, -0.108, 0.496) = +0.86.

In other words, an analyst or decision-maker working with these two experts has very little confidence that in future assignments their estimates would be close enough to be of any use without additional information (such as from a third expert). Moreover, the regression line (green) suggests that in less populous sub-districts the satellite-based estimates tend to be larger. For sub-districts with combined estimates of more than 10⁴ = 10,000 people, the opposite holds. These findings are valuable if the analyst has the opportunity to confront the experts with the apparent biases and motivate them to make informed revisions.

The limits-of-agreement method, in theory, can be applied repeatedly, in order to investigate the concordance between all pairs among three or more experts. However, as we emphasized, the simplified Beroggi-Wallace method is an option chiefly where the analysts have very limited or no opportunity to discuss the pattern of estimates, let alone all individual estimates, with the experts. Intensive interactions with all participating experts, followed by one or more rounds of revisions, are not part of this scenario.

[Sidebar:] Technicalities of the simplified Beroggi-Wallace method

The original methodology is quite involved (Beroggi and Wallace 2000, op.cit.). It covers elicitation and aggregation over several rounds in which the experts are confronted with the “results of their assessments relative to the other experts” (p. 34) and subsequently may adjust their estimates. These are some technicalities of a simplified version, which covers only one round between experts and analysts. The essential common ground with the original is the evaluation of estimates on two criteria – agreement and consistency:

Suppose n experts judge m objects. They provide estimates of a variable \( X \) of at least interval level, \( x_{ij} \), with \( i = 1, 2, \ldots, m \), and \( j = 1, 2, \ldots, n \). Each object occupies a row, each expert a column, in the table \( X = [x_{ij}] \). The \( x_{ij} \) are to be weighted by weights \( w_{ij} \). For every given object \( i = k \), a combined estimate \( \hat{x}_k \) is sought as the weighted mean of the \( x_{kj} \). The weights are proportionate to the product of two scores, the agreement score and the consistency score. \(^{21}\)

Every estimate \( x_{ij} \) is assigned an agreement score \( a_{ij} \). The score expresses how close, for a given object \( i = k \), \( x_{ij} \) is to the mean estimate \( \bar{x}_k \), or, in other words, how much it expresses the experts’ potential consensus on this object. These scores are calculated separately by each object \( i = k \). It is one minus the ratio of the absolute value of the difference of \( j \)'s estimate \( x_{ij} \) from the group mean \( \bar{x}_k \) to the maximal absolute difference among all experts. For example, if experts A, B, C and D submit 10, 20, 30, 100 as their estimates of \( X \) in object \( k \), the group mean is \( 160/4 = 40 \), and the largest absolute difference is \( |100 - 40| = 60 \). The agreement score for object \( k \) and expert A thus is \( a_{k1} = 1 - |10 - 40| / |100 - 40| = 1 - 30/60 = 0.5 \). For D, the outlier, it is \( a_{k4} = 0 \). The agreement score table \( A = [a_{ij}] \) has the same dimensions as \( X \).

\(^{21}\) We use the cumbersome expression “For every given object \( i = k, .. \)” to make sure the reader understands that the calculation runs only within the specific object \( k \), not over all the \( i \)'s.
The **consistency score** is a property of a particular expert \( j = l \). It is not a property of the combination of object and expert, in contrast to the agreement score. However, it is determined by the position of \( l \)'s estimates on all objects relative to those of the other experts. To calculate the score, we go through several steps:

- **First**, we rank the estimates \( x_{kj} \) within each given object \( i = k \), with the lowest \( x_{kj} \) assigned rank = 1 and the highest rank = \( n \). In other words, we rank **every row** of estimates separately. The rank table \( \mathbf{R} = [r_{ij}] \) has the same dimensions as \( \mathbf{X} \).

- **Second**, we consider the ranks of the particular expert \( l \) on all objects \( i \). In other words, we consider all the ranks \( r_{il} \) down the same **column** \( l \) in the rank table. If the number of experts, \( n \), is even, an expert would be maximally inconsistent if her estimates were the lowest (rank = 1) for half the objects, and the highest (rank = \( m \)) for the other half. The mean rank of her judgments would be \( (n + 1) / 2 \), and her mean absolute deviation from her mean rank would be \( (n - 1) / 2 \) [because\(^{22}\)]. Conversely, an expert whose estimates held the same rank on every object would be maximally consistent. She would have a mean deviation of zero. For odd \( n \), the deduction of the mean absolute deviation is less intuitive, but the difference is not fundamental.

- **Third**, taking into account those two extremes, some reasonable consistency measure must be defined. Let \( \text{MAD}_l \) be an expert \( l \)'s mean absolute deviation from her own mean rank, defined as \( \text{MAD}_l = (\sum_i |r_{il} - \bar{r}_l|) / m \). One way to define her consistency score will then be

\[
C_l = 1 - \frac{\text{MAD}_l}{(\text{maximum possible mean absolute deviation from the expert's mean rank})} = 1 - \frac{\text{MAD}_l}{(n - 1) / 2} = 1 - 2\text{MAD}_l / (n - 1).
\]

The table of consistency scores \( \mathbf{C} = [c_j] \) has dimensions 1 x \( n \) (every one of the \( n \) experts is assigned exactly one score)\(^{23}\).

One may assume that in practical life, maximally inconsistent experts are a rare species. Thus, except for the rarissime maximally inconsistent expert, nobody will have a zero consistency score.

Finally, all the weights are calculated as \( w_{ij} = a_{ij} \cdot c_j \cdot s_i \), where the \( s_i \) are scaling factors such that for every object \( i \), \( \sum_j w_{ij} = 1 \). The combined estimate of \( X \) on object \( i \) is

\[
\text{interrater agreement, defined as } AD_{M(j)} = \frac{\sum_{k=1}^{N} |x_{jk} - \bar{x}_j|}{N},
\]

"where \( N \) is the number of judges, or observations, of item \( j \), \( x_{jk} \) is equal to the kth judge's rating of item \( j \), and \( \bar{x}_j \) is equal to the mean rating of item \( j \)" (Smith-Crowe, Burke et al. 2013:132). As the formula shows, the \( AD_{M(j)} \) is defined for an item, or object in our terminology, and is computed by **row**. Our \( \text{MAD}_l \) is a measure of the rater or judge, or expert in our terminology, and is calculated with the values (ranks) by **column**. The extensive discussion of the theoretical and methodological bases of the \( AD_{M(j)} \) in op.cit. is not relevant for our purposes.

\(^{22}\) For the half of his judgments at the lowest rank, the sum of absolute deviations is \( = (m/2) \cdot |1 - (n+1)/2| = (m/2) \cdot (n/2 - 1/2) \). For the half at the highest rank, the sum is \( = (m/2) \cdot |n - (n+1)/2| \), which is also \( = (m/2) \cdot (n/2 - 1/2) \). Thus the sum over all judgments is \( m \cdot (n/2 - 1/2) \). For the mean, divide by the \( m \) objects: \( (m/2) \cdot (n/2 - 1/2) = (n/2 - 1/2) = (n - 1) / 2 \).

\(^{23}\) Readers familiar with the concept of mean absolute deviation in the context of interrater agreement may feel some confusion at this point. They may have worked with the average deviation as an index of
As noted before, an administrative aggregation, if desired, then consists in the simple sum over all objects of interest, \( \sum \hat{x}_i \).

We demonstrate this with fictitious estimates on 6 objects by 4 experts. For visual simplification, we assume that the true value of \( X \) in every one of the six objects is \( = 10 \). Since both the ranks of the estimates and the agreement scores are calculated for each object independently from the estimates of all other objects, this artificial scenario does not entail a loss of generality.

Our experts’ estimates follow different patterns. Expert A, as seen in the raw estimates table below, estimates \( x_i = 9 \) for all six objects – she has a slight, but constant downward bias. Expert B’s estimates are 8 and 12, each three times, with an accurate mean, but little precision. Expert C’s bias is upward and larger; like A’s, it is constant. Expert D is the most interesting participant. Four times she is right on target. On objects 5 and 6, however, her estimates are extremely large. An obvious question is how D’s estimates impact the combined estimates. We compare results using the simplified Beroggi-Wallace method to those based on the arithmetic mean and on the median.
Aggregation and synthesis - Aggregation of quantitative expert judgments

Table 8: A numeric example of the simplified Beroggi-Wallace method

<table>
<thead>
<tr>
<th>Object</th>
<th>Expert A</th>
<th>Expert B</th>
<th>Expert C</th>
<th>Expert D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>8</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>12</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>8</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>12</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>8</td>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>12</td>
<td>13</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 8: A numeric example of the simplified Beroggi-Wallace method (continued)

<table>
<thead>
<tr>
<th>Object</th>
<th>Agreement scores</th>
<th>Unadjusted weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.67 0.33 0.00 1.00</td>
<td>0.22 0.11 0.00 0.44</td>
</tr>
<tr>
<td>2</td>
<td>0.00 0.50 0.00 0.50</td>
<td>0.00 0.17 0.00 0.22</td>
</tr>
<tr>
<td>3</td>
<td>0.67 0.33 0.00 1.00</td>
<td>0.22 0.11 0.00 0.44</td>
</tr>
<tr>
<td>4</td>
<td>0.00 0.50 0.00 0.50</td>
<td>0.00 0.17 0.00 0.22</td>
</tr>
<tr>
<td>5</td>
<td>0.60 0.53 0.87 0.00</td>
<td>0.20 0.18 0.19 0.00</td>
</tr>
<tr>
<td>6</td>
<td>0.50 0.71 0.79 0.00</td>
<td>0.17 0.24 0.17 0.00</td>
</tr>
</tbody>
</table>

The estimates under the Beroggi-Wallace method are overall closest to the true values. This is so because the agreement score mechanism cancelled any effect of Expert C’s estimates on the combined estimates in four out of six objects. It thus neutralized the C’s consistent upward pull in large degree. It also cancelled the effect of D’s extreme values.

Why is the median as an estimator of the combined values not good enough? In this worked example, the row medians are close to the Beroggi-Wallace results, certainly closer than to the row means. If the difference is not large, the far greater convenience of calculating the medians should prevail. For, in real life we cannot know the true values; all we can do is to compare whether the medians overall are closer to the means (which we know are not robust to extreme outlier estimates) or to the Beroggi-Wallace-based results.

There are other reasons to doubt the benefits of this method. Ties – when experts share a common rank because of identical estimates – are difficult to model. Missing values require adjustments that are particularly tedious in Excel. Also, when the underlying variable likely has a skewed distribution (such as area populations, which tend to be log-normally distributed), it may need transformation, and the combined...
Aggregation and synthesis - Aggregation of quantitative expert judgments

estimates will then need to be converted back to the original scale (this complication
would also apply to the median and mean-based estimators).

The justification of the Beroggi-Wallace method is a substantive one. Each of the
medians takes only the estimates regarding a particular object in account. Beroggi-
Wallace moderate this information with a judgment on the consistency of each expert.
This may be of advantage particularly when the experts judge a great number of
objects, as in our example of over 200 Syrian sub-district populations. In other words,
the analyst is not limited to evaluating the distribution of estimates for each particular
object; the assessment and consistency scores tell him something about the quality of
the experts as well.

Table 9: Indicators of expert quality, Beroggi-Wallace method

<table>
<thead>
<tr>
<th>Expert A</th>
<th>Expert B</th>
<th>Expert C</th>
<th>Expert D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean agreement scores</td>
<td>0.41</td>
<td>0.49</td>
<td>0.28</td>
</tr>
<tr>
<td>Consistency scores</td>
<td>0.33</td>
<td>0.33</td>
<td>0.22</td>
</tr>
<tr>
<td>Relative influence (based on adj. w.)</td>
<td>20%</td>
<td>31%</td>
<td>11%</td>
</tr>
</tbody>
</table>

Thus, when we compare the four experts of our fictitious example, surprisingly A and
B, with the constant estimates wind up as less influential than B and D. The method
apparently is quite forgiving of D’s two extreme estimates. These may alert an
attentive analyst – maybe D does know something special! Neither the median nor the
mean method of combined estimates would tell us that!

Multiple-expert estimates of a proportion

Experts may be asked to produce estimates of the proportion of a trait or event of
interest. They may not have estimates of the relevant numerators and denominators for
the particular population for which the analyst or decision-maker wishes to obtain the
proportion in point. Or they may have point estimates, but no measure of their
uncertainty. They may, however, be able to extrapolate from estimates that they formed
about populations that they consider comparable, or to arrive at an estimated proportion
via some other model calculations. For example, an expert may form a rough estimate
of the current proportion of displaced persons in a district by evaluating the relief
committee figures in several recent months, together with an assumption of what
proportion of new arrivals are likely not yet registered. Another expert, asked about the
same district, may use changes in medical clinic attendance, to form her own estimate
of the proportion.

The analyst, receiving proportion estimates from several experts, wishes to combine
them in one estimate.

We distinguish two situations:

- Each of the experts specifies on how many distinct independent observations
  she bases her estimate.
The experts submit estimates, but not the number of observations on which they base them. The analyst may, or may not, have information that justifies weighting the experts differently.

**Estimated proportion, with known observation base**
In the elicitation chapter, we discuss a fictitious scenario of four experts estimating the proportion of villages with disrupted water supplies. The observation basis for each expert is the number of villages in which she stopped on her recent circuit through the district. In a set-up different from pure survey data, we assume that at every stop the experts discussed the situations of the local as well as of an unreported number of surrounding communities.

In this case the point estimate and the confidence interval are obtained via the beta distribution. We refer the reader to the worked example on pages 49 - 52 (calculation relies solely on Excel functions). The challenge is rather in the elicitation than in the aggregation; thus further discussion at this point is not needed.

**Estimated proportion, with unknown observation base**
When the number of independent observations on which the experts base their estimates is not known – they may themselves not be able to enumerate them –, at most the analyst will have other information that allow him to weight the experts. In other cases, even this differential treatment may not be possible or is rejected on policy or other grounds. In this study, we generally refrain from calibrating or otherwise weighting expert judgments. In the situation at hand, we will walk the reader through a formula meant to do the weighting and then demonstrate an example with equal-weighted, i.e. unweighted, experts.

We approach the problem using Bordley’s “*multiplicative formula for aggregating probability assessments*” (Bordley 1982). We banish the technicalities to a sidebar, then continue with a less technical rationale and with a worked example using IDP estimates from Aleppo, Syria.

---

**Sidebar:] Technicalities of Bordley’s formula**
We rely on the presentation in Allard et al.’s (2012) “Probability Aggregation Methods in Geoscience”. Bordley (and others) treat estimated proportions the same way as probabilities. Thus, the expert judgment that “in this district, I believe, 80 percent of the population are IDPs” is equivalent to “the probability that a randomly selected individual from this population is a displaced person is 0.8”. His formula requires the user to work with an additional concept – the *odds* of a proportion. We take this minor detour in order to demonstrate:

- The aggregated value under Bordley’s formula differs from the arithmetic and geometric mean, as well as from the median, of the experts’ estimates. Bordley’s is sensitive in particular to high proportions that one or a minority of experts submit.
- If there is a common belief in a particular proportion *before* the experts contribute their estimates, this prior belief can be taken into account with a weight that the analyst assigns, based on the strength of the prior evidence or
Aggregation and synthesis - Aggregation of quantitative expert judgments

of his own belief (or that of the decision-maker or other powerful stakeholders). If no such prior belief exists, the formula further simplifies.

- If the experts are not weighted, the formula simplifies even further. In this reduced form, we demonstrate an example with MS Excel formulas (the full Bordley’s formula complexity is perfectly manageable in a spreadsheet, but would not fit on this page).

In everyday language, and particular in the world of gambling, the probability of an event happening is often expressed as the “odds for” or the “odds of” [e.g., this or that horse winning]. Mathematically, the odds of a binary event A are expressed as $O(A) = \frac{p(A)}{1 - p(A)}$, the ratio of the probability that A happens to the probability that it does not. In the inverse direction, one obtains the probability from given odds as $p(A) = \frac{O(A)}{1 + O(A)}$ (Wikipedia 2016f). In the example of “in this district, 80 percent of the population are IDPs”, the odds of a person to be an IDP are $O(IDP) = 0.8 / (1 - 0.8) = 4.0$, or in everyday language: four to one.

Bordley demonstrated that the only aggregation operator that satisfied certain mathematical axioms involved the product of the odds of all the expert’s estimated proportions. As noted, the proportions can be weighted; also, they can be moderated with a prior common belief. We explain the full formula in Allard et al.’s (op.cit, 557) notation, then simplify it.

$$O_G(A) = O_0(A)^{w_0} \prod_{i=1}^{n} \left(\frac{O_i(A)}{O_0(A)}\right)^{w_i}$$

where $\prod$ is the product operator. $O_G(A)$ is the aggregated estimate of the probabilities of event A (the subscript G stands for “Global”) or, in our application, the aggregated value of the proportions estimated by experts $i = 1, 2, .. n$. $O_i(A)$ stands for the odds of the proportion believed prior to the expert judgments; the $O(A)$, $i = 1 .. n$, are the odds of the experts estimates. The $w$’s are the weights; they affect the proportions as exponents. Bordley allows all the weights to take any values from zero to +infinite. Our intuition is more restrained: an analyst evaluating expert beliefs will plausibly set weights such that they sum to one – it is hard, in the context of expert judgment, to grasp the meaning of weights that “add up to more than 100 percent”.

Now we simplify. If prior beliefs do not matter, then $O_0(A) = 1$ and $w_0 = 0$. This is so because event A, prior to the expert judgments, and its opposite, not-A, are equally likely, thus $p_0(A) = 0.5$, and the odds, by their definition above, are $O_0(A) = 0.5 / (1 - 0.5) = 1$. $w_0 = 0$ is math speak for “the prior belief does not matter”. Bordley’s formula shrinks to:

$$O_G(A) = \prod_{i=1}^{n} O_i(A)^{w_i}, \text{ with } \sum_{i} w_i = 1.$$  

Moreover, if the analyst does not weight the experts, this is the same as saying that all the $n$ experts have equal weights, $w_i = 1/n$, in which case the formula further simplifies to

$$O_G(A) = \prod_{i=1}^{n} O_i(A)^{1/n}$$

The aggregated proportion then is recovered as $P_G(A) = O_G(A) / [1 + O_G(A)]$. We will now translate this to a simple spreadsheet example.
We revert to our four experts estimating the proportions of village communities with disrupted water supplies. This time the analyst is without the benefit of known observation bases. At first, we use the same estimates as before. The table shows how Bordley’s formula can be implemented in Excel, the resulting aggregated proportion as well as the alternative values produced with other, simpler estimators:

Table 10: Spreadsheet implementation of the simplified Bordley’s formula

<table>
<thead>
<tr>
<th>Expert</th>
<th>Estimated proportion</th>
<th>Odds</th>
<th>odds*(1/n)</th>
<th>Product of all the (odds*(1/n))</th>
<th>Aggregated proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td>0.333</td>
<td>0.760</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.50</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.50</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.80</td>
<td>4.000</td>
<td>1.414</td>
<td>1.075</td>
<td>0.518</td>
</tr>
</tbody>
</table>

Formulas:
- \( \text{Product} = \frac{\text{Formula} - 1}{1 - \text{Formula}} \)
- \( \text{Formula} = \frac{\text{Product}}{\text{Product} - 1} \)

Comparison with other estimators:
- Geometric mean: \( \text{GEOMEAN}([A; B; C; D]) \)
- Arithmetic mean: \( \text{AVERAGE}([A; B; C; D]) \)
- Median: \( \text{MEDIAN}([A; B; C; D]) \)

The differences are minor. Given these particular estimates, they are not worth the conceptual and mechanical effort to calculate Bordley’s variant.

Now, however, consider the case that expert D feels she has the information justifying an extreme estimate – 95 percent of all villages are living with disrupted water supplies. Although the estimate is extreme, the analyst is taking it seriously; he assigns D the same weight as A, B and C. We find:

Table 11: An example with one expert providing a very high estimate

<table>
<thead>
<tr>
<th>Expert</th>
<th>Estimated proportion</th>
<th>Odds</th>
<th>odds*(1/n)</th>
<th>Product of all the (odds*(1/n))</th>
<th>Aggregated proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td>0.333</td>
<td>0.760</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.50</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.50</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.95</td>
<td>19.000</td>
<td>2.088</td>
<td>1.586</td>
<td>0.613</td>
</tr>
</tbody>
</table>

Formulas:
- \( \text{Product} = \frac{\text{Formula} - 1}{1 - \text{Formula}} \)
- \( \text{Formula} = \frac{\text{Product}}{\text{Product} - 1} \)

Comparison with other estimators:
- Geometric mean: \( \text{GEOMEAN}([A; B; C; D]) \)
- Arithmetic mean: \( \text{AVERAGE}([A; B; C; D]) \)
- Median: \( \text{MEDIAN}([A; B; C; D]) \)

The two flavors of the mean have barely responded. The median is entirely robust, as it should be. Bordley’s formula, however, produced an increase of almost 0.10. It is more sensitive to probabilities towards the high extreme; if D had made a case for 0.975, the aggregated estimate would jump to 0.655.
Aggregation and synthesis - Aggregation of quantitative expert judgments

A few general points can now be made; the reader may already have anticipated them:

- Bordley’s multiplicative formula is suitable only if none of the experts provides estimates at the theoretical extremes – probabilities of strictly zero or one. Anyway, that would be rare - if the trait or event of interest were known with certainty, decision-makers would not turn to experts.
- It is attractive when the lowest and highest estimates enjoy the same confidence, and thus carry the same weight, as the inliers. If the former are distrusted and weighted downwards, then the median of the proportion will seem more reassuring. However, the analyst should not give lower weights simply because the estimate is an outlier, but, if at all, because the supporting evidence for the estimate is weaker.
- In many situations, Bordley’s and the geometric mean results will be close to each other. The geometric mean, however, does not take into account that the same relative change in estimated proportions towards the higher extreme, e.g., from 0.90 to 0.99, is more dramatic than at the center of the scale, e.g. from 0.50 to 0.55. Bordley’s formula, based on the odds, is sensitive to this.

More importantly for the actual business of humanitarian expertise, Bordley’s formula is indifferent to the sequence in which the experts arrive. As we emphasized earlier, for efficiency and cost reasons, it often makes sense to add experts sequentially. Suppose that the decision-maker initially hires three experts A, B and C. They all estimate that the trait of interest is present in 60 percent of the population, \( p = 0.6 \). Obviously, all four estimators that we discussed produce an aggregated estimate of \( p_{G,3\text{experts}} = 0.6 \). Suspicious of the experts’ unanimity, the decision-maker commissions a fourth expert, D. Suppose this expert D comes up with a radically deviant estimate of 0.99. The analyst now is tasked to compute the updated aggregated estimate.

Should she combine the unanimous opinion of 0.6 as the first estimate, to be combined with 0.99 as the later, with equal weights? Or should she consider A, B, C and D’s estimates as individual and simultaneous and combine three times 0.6 and once 0.99?

Using the geometric mean estimator, the results are not the same. The former path leads to \((0.6 \times 0.99)^{\frac{1}{2}} = 0.771\), the latter to \((0.6^3 \times 0.99)^{\frac{1}{4}} = 0.680\). Bordley proved that his formula was the only metric that ensured what he called “pool, then update = update, then pool”. “Pool” stands for “pooling opinions”, here = “aggregating expert estimates”. In this example, Bordley’s result is 0.810 both ways\(^{24}\). This property makes this formula particularly attractive in contexts where experts may become involved sequentially, or when some of the estimates are delayed, but ultimately have to be integrated.

---

\(^{24}\) Readers may want to demonstrate this by themselves in a spreadsheet. Hint: In the instance where all expert estimates are entered simultaneously, the twice simplified Bordley formula, as implemented in the spreadsheet above, will directly furnish the result, \( p_G = 0.8104 \). However, when A, B and C’s estimates are first pooled, and D’s is added later (the sequential case), the full Bordley formula comes into play. The pooled \( p_{G,3\text{experts}} = 0.6 \) becomes the prior belief \( p_0 \) (before expert D joins), with the associated odds \( O_0 = 1.5 \), \( p_0 = 0.99 \), thus \( O_0 = 99.0 \). It is fully admitted, ie. \( w_0 = 1 \). The exponential weight on \((O_D / O_0)\) is 0.25 (D is one of four experts). Thus, according to Bordley, \( O_D = O_0 \times (O_D / O_0)^{0.25} = 1.5 \times (99 / 1.5)^{0.25} = 4.2754 \), and \( p_D = O_D / (1 + O_D) = 4.2754 / 5.2754 = 0.8104 \). Readers interested in the underlying mathematics find it in Allard et al., op.cit.
Aggregation and synthesis - Aggregation of quantitative expert judgments

[Sidebar:] The proportion of IDPs in Aleppo, Syria, summer 2015

OCHA, as part of the data collection for the 2016 Humanitarian Needs Overview in Syria, collected estimates of IDPs and returnees by sub-district during summer 2015.

They came from three sources: the so-called Governorate Profiles (a joint statistical effort between the Government of Syria and the United Nations), the OCHA-led Whole-of-Syria Approach (WoSA), and the Needs and Population Monitoring Project (NPM). Generally, WoSA supplied several key informant estimates from every sub-district, the other two sources one each. Unlike the Governorate Profile and WoSA sub-district population estimates, which came also with minima and maxima (see the aggregation of triangular distributions, pages 68-74), regarding IDPs and returnees the sources supplied only best estimates. For IDPs, 1,704 estimates were usable, covering 267 sub-districts. The 1,506 usable returnee estimates covered 196 sub-districts.

As there were multiple estimates for most of the sub-districts, some of them covering only parts of the sub-district area, there was a need to reduce them to one per sub-district. At the time, this was done by calculating the proportions of IDPs and returnees for each initial estimate, and then weighting them by their estimated sub-district population size. For sub-districts with missing values, the median proportions across the observed sub-districts were substituted (24.3 percent for the IDPs; 3.1 percent for the returnees). These unified proportions were then multiplied by the sub-district populations that had been computed separately, by integrating Governorate Profiles, WoSA as well as satellite imagery-based estimates.

In hindsight, that may not have been the optimal procedure. There is no reason why an estimate of the IDP proportion, coming from a key informant estimating a lower population or speaking for part of the sub-district only, should be farther from the true value. For, even the key informants speaking for entire sub-districts plausibly were basing their judgments on the parts most familiar.

Pooling the estimated proportions with equal weights now seems more appropriate. Thus we want to demonstrate Bordley’s formula for pooling proportions with equal weights and, for completeness, also with variable ones. We do this with the estimates of IDPs in one sub-district, Jebel Saman in Aleppo. Besides the one estimate in the Governorate Profiles and that of NPM, WoSA added estimates from 12 of its key informants. These are the data, ordered by source and then by population size.
Aggregation and synthesis - Aggregation of quantitative expert judgments

Table 12: Population and IDP estimates for a sub-district in Syria, 2015

<table>
<thead>
<tr>
<th>Source</th>
<th>Population</th>
<th>Estimated IDPs</th>
<th>IDP proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gov.Prof.</td>
<td>1,200,000</td>
<td>508,000</td>
<td>42%</td>
</tr>
<tr>
<td>NPM</td>
<td>2,927,295</td>
<td>770,210</td>
<td>26%</td>
</tr>
<tr>
<td>WoSA</td>
<td>360,000</td>
<td>135,000</td>
<td>38%</td>
</tr>
<tr>
<td>WoSA</td>
<td>380,000</td>
<td>130,000</td>
<td>34%</td>
</tr>
<tr>
<td>WoSA</td>
<td>390,000</td>
<td>60,000</td>
<td>15%</td>
</tr>
<tr>
<td>WoSA</td>
<td>400,000</td>
<td>110,000</td>
<td>28%</td>
</tr>
<tr>
<td>WoSA</td>
<td>450,000</td>
<td>175,000</td>
<td>39%</td>
</tr>
<tr>
<td>WoSA</td>
<td>500,000</td>
<td>300,000</td>
<td>60%</td>
</tr>
<tr>
<td>WoSA</td>
<td>2,036,000</td>
<td>1,196,000</td>
<td>59%</td>
</tr>
<tr>
<td>WoSA</td>
<td>2,220,000</td>
<td>1,152,000</td>
<td>52%</td>
</tr>
<tr>
<td>WoSA</td>
<td>2,300,000</td>
<td>1,700,000</td>
<td>74%</td>
</tr>
<tr>
<td>WoSA</td>
<td>2,300,000</td>
<td>1,700,000</td>
<td>74%</td>
</tr>
<tr>
<td>WoSA</td>
<td>3,350,000</td>
<td>1,050,000</td>
<td>31%</td>
</tr>
<tr>
<td>WoSA</td>
<td>3,400,000</td>
<td>1,060,000</td>
<td>31%</td>
</tr>
</tbody>
</table>

The population-weighted mean IDP proportion works out as 46.1 percent. The individual estimates vary greatly, from a low 15 percent to a high 74 percent. The estimated proportion is weakly positively correlated with the population denominator (Spearman’s rho = 0.13). First, we treat all estimates with equal confidence, i.e. with equal weights $w_i = 1/14$. Second, we consider NPM’s reputation, at the time, for more highly disaggregated and better ground-truthed estimates. An analyst placing higher confidence in this source might allocate different weights, such as $w_{NPM} = 1/3$ and $w_{GP} = w_{any\_WoSA} = 2/3 \times 1/13 = 2/39$. Together these weights sum to 1.
Table 13: Aggregation of estimated IDP proportions

<table>
<thead>
<tr>
<th>Source</th>
<th>Estim. IDP proportion</th>
<th>Odds</th>
<th>Equal weights</th>
<th>NPM higher weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Weight</td>
<td>Odds*Weight</td>
</tr>
<tr>
<td>Gov.Prof.</td>
<td>42.3%</td>
<td>0.7341</td>
<td>0.0714</td>
<td>0.9782</td>
</tr>
<tr>
<td>NPM</td>
<td>26.3%</td>
<td>0.3571</td>
<td>0.0714</td>
<td>0.9291</td>
</tr>
<tr>
<td>WoSA</td>
<td>37.5%</td>
<td>0.6000</td>
<td>0.0714</td>
<td>0.9642</td>
</tr>
<tr>
<td>WoSA</td>
<td>34.2%</td>
<td>0.5200</td>
<td>0.0714</td>
<td>0.9544</td>
</tr>
<tr>
<td>WoSA</td>
<td>15.4%</td>
<td>0.1818</td>
<td>0.0714</td>
<td>0.8854</td>
</tr>
<tr>
<td>WoSA</td>
<td>27.5%</td>
<td>0.3793</td>
<td>0.0714</td>
<td>0.9331</td>
</tr>
<tr>
<td>WoSA</td>
<td>38.9%</td>
<td>0.6364</td>
<td>0.0714</td>
<td>0.9682</td>
</tr>
<tr>
<td>WoSA</td>
<td>60.0%</td>
<td>1.5000</td>
<td>0.0714</td>
<td>1.0294</td>
</tr>
<tr>
<td>WoSA</td>
<td>58.7%</td>
<td>1.4238</td>
<td>0.0714</td>
<td>1.0256</td>
</tr>
<tr>
<td>WoSA</td>
<td>51.9%</td>
<td>1.0787</td>
<td>0.0714</td>
<td>1.0054</td>
</tr>
<tr>
<td>WoSA</td>
<td>73.9%</td>
<td>2.8333</td>
<td>0.0714</td>
<td>1.0772</td>
</tr>
<tr>
<td>WoSA</td>
<td>73.9%</td>
<td>2.8333</td>
<td>0.0714</td>
<td>1.0772</td>
</tr>
<tr>
<td>WoSA</td>
<td>31.3%</td>
<td>0.4565</td>
<td>0.0714</td>
<td>0.9455</td>
</tr>
<tr>
<td>WoSA</td>
<td>31.2%</td>
<td>0.4530</td>
<td>0.0714</td>
<td>0.9450</td>
</tr>
</tbody>
</table>

Product of the weighted odds: 0.7362 0.6003
Aggregated proportion: 42.4% 37.5%

Comparison estimators:
- Geometric mean: Unweighted: 39.6% Weighted: 35.3%
- Arithmetic mean: Unweighted: 43.1% Weighted: 38.4%
- Median Unweighted: 38.2% Weighted: 31.3%

With the estimates all way inside the [0, 1] theoretical range of probabilities, the differences in the unweighted case are minor; any of the alternative methods would be good enough. However, in Excel, weighted means and medians are more cumbersome to calculate than Bordley’s formula (for convenience, we calculated those values in the statistical application Stata). This confirms the insight that fast and simple rules (e.g., “Take the median of the proportions”) will not work well as soon as additional requirements come into play or the values of key variables are outside typical ranges.

Despite the elegance of Bordley’s formula, the result is unsatisfactory. The decision-maker may be happy with just one number. But, surely, the analyst will want to have some measure of uncertainty. Is there a rational measure? We know that there were 14 estimates, and at best they were independent one from the other. In the extreme, we can argue that the estimates arrived from the same probability distribution as if each of the key informants had randomly pulled one person from Jebel Saman’s population, together producing a sample of 14 persons. Let us assume that the true proportion of IDPs in Jebel Saman was equal to Bordley’s result (unweighted), i.e. 42.4 percent. Thus a key informant randomly pulling out one person has a 42.4 percent chance that she found an IDP. To see the probabilities that a sample of 14, with success probability of 0.424, will contain at least X IDPs, Excel’s BINOM.DIST comes to our help. We seek the number of successes with close to 0.025 and 0.097 with the aim to form an approximate 95-percent confidence interval. By tabulating all possible s = 1, .., 14, we find that BINOM.DIST(x=2, n=14, p=0.424, TRUE) = 0.0268, and BINOM.DIST(9,14,0.424,TRUE) = 0.9726. Thus one can, on that reasoning, suggest a confidence interval of [2/14, 9/14] = [14, 64] percent IDPs in Jebel Saman.
Surprisingly, this is narrower than the range of the 14 estimates, which goes all the way from 15.4 to 73.9 percent.

This method of estimating a confidence interval for estimated proportions is tempting\textsuperscript{25}. Yet it is not valid – the key informants would not agree that their combined estimates have the same information value as a sample of 14 randomly drawn persons. The conclusion is: Yes, there is a best estimate, but there is no rational confidence interval around it. The information base of the individual key informants remains unknown. The equal weights for all of them express the analyst’s lack of knowledge, not equal sampling probabilities. If the decision-maker doesn’t like the best estimate, he/she will have to make the case for higher weights for some and lower weights for others, or insist on a high weight for his/her personal estimate as the prior belief.

4.4. **Expert judgment and Bayesian reasoning**

The philosophy of humanitarian assessments holds that impacts and needs become known through a process of successively refined measurement and estimation. Thus, in the first few weeks after a sudden-onset disaster, the information collection and analysis move towards more quantitative information, sharper sectoral perspectives, finer granularity and better statistical representation.

This claim, however, is not universally valid. First, it is not only measurements that are expected to improve, but also the process models that combine them in order to produce key measures such as the severity of unmet needs. Second, qualitatively, the things that are important to know change over time, such as from needs assessments to response plans. Third, the substitution of detailed and current survey data for broad expert opinion or extrapolated secondary information is not one-way. To exemplify, the most detailed survey of the needs for farm inputs makes sense only in the light of the seasonal calendar. One competent expert may be enough to judge the flexibility of the calendar and the chances of relief crops sown after a certain date.

Nonetheless, interest in some variables will likely persist. Thus, figures of people with access to safe drinking water, for example, are broadly estimated in the emergency, then more closely monitored during reconstruction. Periodic updates counteract the natural obsolescence of the information.

The idea of updating knowledge on the strength of fresh data is central to a branch of statistics known as Bayesian analysis. It takes its name from Thomas Bayes (1701 – 1761) (Wikipedia 2017\textsuperscript{k}), an English Presbyterian minister credited with the formulation of its central theorem. This section offers a few glimpses at Bayesian thinking in or close to expert judgment; apart from stating and exemplifying Bayes’ theorem, deeper statistical expositions are not intended.

There are several reasons why humanitarian analysts should take at least superficial note of Bayesian reasoning, beyond the shared concern for updated knowledge:

\textsuperscript{25} Similarly, analysts might be tempted to bootstrap mean and confidence interval. This presumes that the key informants were sampled, with equal probability, from a population of – who? – key informants!
While Bayes’ ideas remained nearly dormant for a long time, in the last fifty years modern statistics has been shaken by a Bayesian revolution. The revolution is increasingly building strongholds in fields adjacent to expert judgment, such as in risk analysis and evaluation theory. Its momentum is, if anything, growing.

Some researchers (e.g., Humphreys and Jacobs 2015) claim that Bayesian methods can be employed to better integrate qualitative and quantitative information. Their reasoning style has inspired adherents of “process tracing” (Befani and Stedman-Bryce 2016), a causal-analysis method that is of interest also to students of expert judgment.

In the field of expert judgment, analysts must respect that decision-makers often consider themselves experts. Thus, the decision-makers’ beliefs prior to the experts’ work matter. The aggregated estimate is a weighted average of the decision-maker’s personal (or his/her agency’s official) and the experts’ professional estimates. A perspective like the Bayesian that forces consideration of prior assumptions is helpful in such contexts.

We proceed as follows: Bayes’ theorem is the basis of all Bayesian reasoning even if experts pursue a train of Bayesian thought with simple intuitive, non-formal means. We present the formula, some notation and the simple geometric proof that the Wikipedia article offers. We encourage the reader to work through this; the mathematics is not overwhelming; the theorem is so fundamental that it underpins much more than a broad conviction that “updating is important”.

Then, we demonstrate how analysts can update their beliefs on the strength of new evidence (including evidence supplied by experts). The dynamic allocation of needs assessment resources exemplifies Bayesian thinking without the numbers game. A visual method borrowed from medical testing gives a rationale for updating probability ranges. Finally, we discuss process tracing as a method of investigating causal claims, with an illustration from a cholera epidemic in Haiti.

**Bayes’ theorem**

We start with a fictitious example from the humanitarian world:

In a war-torn country, violent events have taken a toll on piped water supplies in some residential areas of a city. In recent months, increasing numbers of IDPs have arrived; most of the families are living in cheap rented spaces. The Relief Committee estimates that a third of the residential units are occupied by displaced families. The pattern of functioning water meters leads the Utilities Dept. engineers to estimate that 55.6 percent of the units still receive piped water (only engineers can be this precise!). To assess the technical nature of the damage, the engineers conduct a small sample survey of units with broken supplies. They find that 50 percent of the sample units are occupied by displaced families.

---

26 McGrayne (McGrayne 2011) provides an engrossing, if slightly sensationalist, history of over 200 years’ worth of stagnation and progress of Bayesian thinking, with an appendix demonstrating several simple applications of Bayes' theorem.
The Relief Committee wonders whether these pieces can be connected in a way to estimate the proportion of IDP households with piped water supplies.

To answer the question, we turn to Bayes’ theorem. We introduce it in a restricted didactic version first and later add the more general interpretation. The theorem connects the probabilities that members of a population have two attributes, A and B. Specifically, it formulates a rule about the probability that a member has attribute A if it is known that he/she has B. The probability of having attribute A is written as $P(A)$; the probability of A when B is true is written as $P(A | B)$ and read as “the probability of A, given B”.

The theorem is written out:

$$ P(A | B) = \frac{P(B | A) \times P(A)}{P(B)} $$

It states that the probability to observe A, given B, is equal to the product of the probabilities to observe B, given A, and to observe A, divided by the probability to observe B.

The Relief Committee wants an estimate of the proportion of IDP families that have running water in their residential units.

We know:

$P(B | A) = P($IDP | broken supplies$) = 0.5$ [from the engineers’ sample survey].

$P(A) = P($broken supplies$) = 1-0.556 = 0.444$ [based on water meters].

$P(B) = P($IDP$) = 1/3$ [Relief Committee estimate].

We obtain:

$P(A | B) = P($broken supplies | IDP$) = P(B|A) \times P(A) / P(B) = 0.5 \times 0.444 / (1/3) = 0.666$. Thus, $P($running | IDP$) = 0.334$. One third of the IDPs living in rented space is estimated to have running water.

The hypothetical city population in this table exemplifies a set-up with these probabilities, as the reader can verify. These figures were, of course, not known to the agencies, who found themselves working with mere estimates and the results of a sample survey.

<table>
<thead>
<tr>
<th>Water supplies</th>
<th>Occupied by</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Residents</td>
<td>IDPs</td>
</tr>
<tr>
<td>Running</td>
<td>10,000</td>
<td>2,500</td>
</tr>
<tr>
<td>Broken</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>15,000</strong></td>
<td><strong>7,500</strong></td>
</tr>
</tbody>
</table>

Why does Bayes’ theorem work? The Wikipedia page on the subject (Wikipedia 2017a) offers a nifty geometric support to a simple proof.
Figure 14: A geometric visualization of Bayes' theorem

Geometrically it is demonstrated that $w / (w + x + y + z)$ can be expressed in two ways. It follows that

$$P(A | B) * P(B) = P(B | A) * P(A)$$

and hence

$$P(A | B) = P(B | A) * \frac{P(A)}{P(B)}.$$ 

It may help some readers to work with the population figures in the preceding table instead of the $w, x, y$ and $z$. We demonstrate this in the sidebar on the following page.

$\bar{A}$ and $\bar{B}$ ("A-bar", "B-bar") mean, and can be written as, “not-A” and “not-B”.

Two more useful relationships are easy to derive from the geometric approach:

$$w / (w + x + y + z) = P(A \land B)$$ [the probability of observing $A$ and $B$]. Thus:

$$P(A \land B) = P(A | B) * P(B), \text{ hence } P(A | B) = P(A \land B) / P(B).$$ The conditional probability of observing $A$, given $B$, equals the probability of observing $A$ and $B$, divided by the probability of observing $B$. Analogously:

$$P(A \land B) = P(B | A) * P(A), \text{ hence } P(B | A) = P(A \land B) / P(A).$$

[Sidebar:] Numeric demonstration of Bayes theorem

For greater intuition, we demonstrate the geometric proof of the theorem in Figure 14 using the population assumptions that we made in Table 14 just above it. This table reproduces the same figures prefaced with the algebraic names.

---

27Attribution: By Cmglee - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=33268427
Table 15: Hypothetical population figures with algebraic variable names

<table>
<thead>
<tr>
<th>Water supplies</th>
<th>Occupied by</th>
<th>Case B:</th>
<th>not-B:</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Residents</td>
<td></td>
<td>IDPs</td>
<td></td>
</tr>
<tr>
<td>Condition A:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Running water</td>
<td>w = 10,000</td>
<td>x = 2,500</td>
<td>w + x = 12,500</td>
<td></td>
</tr>
<tr>
<td>not-A:</td>
<td>y = 5,000</td>
<td>z = 5,000</td>
<td>y + z = 10,000</td>
<td></td>
</tr>
<tr>
<td>Broken supply</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>w + y = 15,000</td>
<td>x + z = 7,500</td>
<td>w + x + y + z = 22,500</td>
<td></td>
</tr>
</tbody>
</table>

The proof in Figure 14 hinged on the fact that the probability to be a resident household AND to have running water, \( P(A \land B) = \frac{w}{w + x + y + z} = \frac{10,000}{22,500} \approx 0.44 \), can be calculated in two ways.

**The geometric proof**

First calculation, in “vertical direction”:

\[
P(A|B) \times P(B) = \frac{w}{w+y} \times \frac{w+y}{w+x+y+z} = \frac{w}{w+x+y+z}
\]

\[
= \left(\frac{10,000}{15,000}\right) \times \left(\frac{15,000}{22,500}\right) = \frac{10,000}{22,500} = 0.44
\]

Second calculation, in “horizontal direction”:

\[
P(B|A) \times P(A) = \frac{w}{w+x} \times \frac{w+x}{w+x+y+z} = \frac{w}{w+x+y+z}
\]

\[
= \left(\frac{10,000}{12,500}\right) \times \left(\frac{12,500}{22,500}\right) = \frac{10,000}{22,500} = 0.44
\]

**Back to algebra and numbers**

By flipping and linking the equations, we obtain

\[
w / (w + x + y + z) = P(A | B) \times P(B) = P(B | A) \times P(A)
\]

or in the figures of the example:

\[
0.44 \approx (10,000 / 15,000) \times (15,000 / 22,500) = (10,000 / 12,500) \times (12,500 / 22,500)
\]

or

\[
0.44 \approx 0.667 \times 0.667 = 0.8 \times 0.556
\]
Dividing both sides by \( P(B) \) gives us the final form of the theorem:

\[
P(A | B) = \frac{P(B | A) P(A)}{P(B)}
\]

and in the example:

\[
10,000 / 15,000 = [(10,000 / 12,500) * (12,500 / 22,500)] / (15,000 / 22,500),
\]

which reduces to \( 10,000 / 15,000 \approx 0.667 \)

**Nice! - But the Relief Committee only knows the probabilities**

The Relief Committee of our motivating example, however, does not know the population counts. It only has relative frequencies at hand:

\[
P(IDP | broken supplies) = 0.5 \text{ [from the engineers’ sample survey]}
\]
\[
P(broken supplies) = 0.444 \text{ [based on water meters]}
\]
\[
P(IDP) = 1/3 \text{ [the Committee’s own estimate]}
\]

The Committee wants an estimate in the “opposite direction”: \( P(broken supplies | IDP) \). By applying Bayes’ theorem, we get

\[
P(broken supplies | IDP) = P(IDP | broken supplies) * P(broken supplies) / P(IDP)
\]

\[
= 0.5 * 0.444 / (1/3) = 0.666. \text{ Two thirds of the IDP families live in residential units with broken water supplies; one third has running water.}
\]

**A generic interpretation of Bayes’ theorem**

The example above connected the probabilities to observe two binary attributes. An object could either be A or non-A, but no third option. Similarly, it could only be B or non-B.

Bayes’ theorem has much wider applications, including in expert judgment, when the meaning of A and B is generalized. While the geometric visualization above illuminates a mere algebraic identity as the basis of the theorem, its Bayesian interpretation goes further. Probability in this tradition expresses a degree of belief. Rewriting the theorem as

\[
P(A | B) = \frac{P(B | A) P(A)}{P(B)}
\]

A and B are now propositions of any kind. The real benefit of this re-interpretation accrues when we consider A as a hypothesis and B as evidence. Thus,

- \( P(A) \) is the initial degree of belief that hypothesis A is true, before finding evidence B. \( P(A) \) is therefore known as the prior.
- \( P(B) \) is the probability of finding evidence B, regardless of whether A is true or not.
- \( P(B | A) \) is the probability of finding B if hypothesis A is true.
- \( P(A | B) \) is the outcome of interest – the probability that hypothesis A is true if evidence B is found. It is known as the posterior.
The ratio $P(B \mid A) / P(B)$ is the multiplier that connects the prior to the posterior. It expresses the support that the evidence provides for the hypothesis. $P(A \mid B)$ is the degree of belief in A, updated with evidence B.

Translated to the world of expert judgment, two situations occur:

- $P(A)$ is the degree of belief with which the decision-maker (and perhaps also the analyst) held an assumption, before consulting experts. B is the essence of the expert judgments (e.g., an aggregated estimate). $P(A \mid B)$ is the degree of belief that the decision-maker still places in A after the experts share their opinions and findings with him. For example, if A stands for “the water in town X is now safe to drink”, and B stands for “experts found one water sample that tested positive for cholera”, with $P(A) = 0.8$ and $P(A \mid B) = 0.1$, the experts’ work turned a good confidence that the water was safe into an acute fear that it might not be safe.

- A itself is a theory entertained by experts, such as the consensus of professionals in a field on a well-established statistical relationship. B is evidence that the decision-maker has their own people collect, and which he/she initially interprets naively or is unable to interpret at all. He/she calls in the experts, who improve upon his/her interpretation, or supply an entirely different one, integrating evidence B with their own theory A. We have already discussed such an example: In the introductory chapter we meet a medical coordinator who relies on tracing teams scouting for newly infected people. The teams can cover a limited number of villages only. Although the teams find zero new patients, the coordinator doubts that the risk of new outbreaks is lower than 2 percent, the bar that he sets to justify winding down services in this area. He contacts an epidemiologist, who has calculated the probability distribution of the risk of new outbreaks from the experience with previous epidemics. She integrates the tracing team findings from 20 villages and advises the coordinator to expand tracing to 20 more.

**Updating beliefs on the strength of new evidence**

The Bayesian perspective is useful also to those who reason chiefly in qualitative terms. Expressing the strength of a belief in terms of probability is necessarily quantitative, but probabilities are assigned to propositions also in everyday language that shades into the qualitative. Saying “We firmly believe that X will happen” assigns a high probability to X, without committing to a number. In theory, experts should be comfortable translating the evidential bases of their judgments into quantitative probabilities. In practice, and particularly in humanitarian situations as yet poorly assessed, qualitative and quantitative measures of the strength of one’s beliefs may be difficult to separate.

The challenge then becomes one of rationally updating one’s beliefs when new evidence comes into play, using both qualitative and quantitative figures of thinking. In this section we illustrate some situations and some instruments that are potentially useful when analysts and experts do not have the data to formally apply Bayes’ theorem.
Allocation of needs assessment resources

A country is hit by a sudden-onset disaster that impacts vast areas, possibly – this is not immediately known – the entire territory. Experts and other outside observers initially know nothing about the distribution of impacts. Since all areas seem equally likely to be severely impacted, initial assumptions are empty. The first panel of this diagram expresses the absence of any specific initial information as a uniform distribution (for simplicity, the country map has been reduced to one dimension, the East-West axis). Bayesians speak of an “uninformative prior”. Rationally, the first assessments are conducted at equal distance across the suspected range of impacted areas (panel 2). Their findings are of severe impacts in the center, tapering off in both directions. This evidence motivates an updated belief, as in panel 3. The new belief constitutes an informative prior to the next round of assessments – no longer is every possibility the same likely as all the others.

Rationally, then, additional assessments are fielded where they promise returning information on highly impacted areas (panel 4). Gaps – because of limited resources and limited access – remain (panel 5); the humanitarian community strives to close them. Unless the additional assessments in panel 4 (red arrows) significantly modify the beliefs about the impact distribution, it is rational to allocate resources for yet more assessments in proportion to the length of the arrows in panel 6.

Practically two sets of beliefs change in parallel. One is about the distribution of impacts. The other is a set of predictions about what additional measurements will yield the most valuable updates about that distribution. The needs assessment community is self-reflective; every update, at least in theory, speaks to object and subject alike.
None of that is particularly new or revealing. The process of information collection can be described without Bayesian vocabulary. Things become trickier when we admit that information grows obsolete unless appropriately updated. Without updating our beliefs with fresh evidence, over time they grow oppressive. Either they become stubborn – refusing to acknowledge that realities change – or revert to overly large uncertainty – assuming that more may have changed than is rational in view of the commonly experienced world (Hubbard 2014:185).

For the practicing humanitarian expert, the more intriguing question is: How much is actually updated, and what intensity of updates is optimal? A general answer does not seem possible, but the documented needs assessment dynamics in a given theater may provide some insight.

Following the two earthquakes that shook Nepal in April and May 2015, killing more than 9,000 people, humanitarian agencies engaged in intensive and sustained assessment activities. The OSSOC/OCHA Assessment Cell in Kathmandu kept track...
of coverage and gaps. Its collection of assessment reports published between April 27 and August 19 eventually grew to 230 documents. The analysis of the coverage, by sector, of the 23 affected districts during that period required some arbitrary definitions. Notably, the shelf-life of assessments was set to three weeks; older information was rated as obsolete. The fraction of districts not covered by at least one recent assessment served as a gap measure.

This figure summarizes the dynamic. In short, the gap in terms of current relevant information was lowest about three weeks after the first quake. It then was allowed to slowly grow; in other words, assessments that by this definition had fallen obsolete were not replaced. The obsolescence of the portfolio accelerated after two months. A bevy of fresh assessments in July narrowed the gap, although not to the level in mid-May. The increase in assessment activity anticipated a new round of funding requests due before many of the initial emergency grants would come to their end.

Figure 16: Nepal - Information gaps over time

![Image of Nepal earthquakes: Information gaps over time graph]

Source: Benini, Chataigner et al., op.cit., 34.

If we consider needs assessment to be the domain of experts, it is plain that the demand for expertise in this case followed pragmatic considerations. The initial effort was massive, so as to give the humanitarian community a secure handle on the situation. Once the relief operations were gathering momentum, the need for continuous updates lessened. A faster pace was resumed on the plausible assumption that the donors to be approached for continued funding would want to see rationales supported with recent assessments of achievements and of continuing need.

**Updating estimated probabilities of an event**

In scenario building exercises as well as in other types of elicitation, experts may advance broad estimates of the probability of a future event. "Between April and September this year, more than 500,000 migrants will cross the Mediterranean. We
Aggregation and synthesis - Expert judgment and Bayesian reasoning

give this event a probability of between 60 and 80 percent” would be a proposition of this kind.

Experts may be more comfortable using a small set of probability categories rather than precise continuous probabilities. Even if they personally have worked out a precise estimate, such as from a simulation, for decision-makers and for public consumption they may want to phrase the result in terms of a range of probabilities. A range reflects the uncertainty better in the eyes of the audience. An update will therefore lead to another proposition asserted within a range of probability – higher, the same, or lower as the prior.

What kind of evidence would it take to update the event probability from a range of 0.6 – 0.8 to, say, 0.8 – 0.9? Surprisingly, the field of medical testing offers some ideas for humanitarian expert judgment. At first sight, this is not plausible – the true nature of a disease is not directly observable; observable symptoms are associated with a number of diseases, each with a certain diagnostic probability. One can stretch this basic situation to latent needs and manifest behaviors in the humanitarian domain, but the parallel grows thin when “symptoms” are replaced with “medical tests”. In needs assessments, there is no equivalent to X-rays. Nutritional surveys are perhaps the closest thing, but there are no generic tests comparable to medicine. “Test relief”, taking food-for-work program participation as a gauge of a gathering food crisis, was a test used by colonial bureaucrats, particularly in India. It is still spooking around in the toolboxes of famine relief and social protection in some countries (Coirolo, Commins et al. 2013), but its meaning and scope are too particular and too compromised by history to nowadays serve as a valid analogue in this context.

We pursue this line of thought for other reasons. The idea of the medical test is that a reliable measurement outcome has a valid connection with a suspected diagnosis. The diagnosis initially may enjoy a low or modest degree of belief. In a strong test, a positive result will move the degree of belief closer to certainty. A negative result will lower belief in the particular diagnosis closer to its exclusion. A strong test, in other terms, is highly sensitive and specific – it causes low rates of false negatives or false positives (Wikipedia 2017j).

A probability scale with unequal intervals

In the context of probability ranges, Medow and Lucey (2011) have taken this further in a Bayesian perspective. Their intuitive ideas are useful at the cusp of qualitative and quantitative expert judgement, at least for their didactic intuition, which relies on visualization.

The authors at first define five probability categories:

Table 16: Medow and Lucey's probability categories

<table>
<thead>
<tr>
<th>Categorical probability</th>
<th>Numerical probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very unlikely</td>
<td>Less likely than 10%</td>
</tr>
<tr>
<td>Unlikely</td>
<td>Between 10% and 33%</td>
</tr>
<tr>
<td>Uncertain</td>
<td>Between 34% and 66%</td>
</tr>
<tr>
<td>Likely</td>
<td>Between 67% and 90%</td>
</tr>
<tr>
<td>Very likely</td>
<td>More likely than 90%</td>
</tr>
</tbody>
</table>
Their basic idea is:

- If a diagnosis is **unlikely, uncertain or likely**, one test is required to move it up or down by one category, depending on the positive or negative test result.
- If a diagnosis is **very unlikely or very likely**, “two tests are needed to escape the very categories” (ibid., p. 166).
- If the **strength of the test** is well known, the updated probability range can be narrowed further.

The figure illustrates how the updated probability range for a likely event is both higher and narrower in response to a positive test result.

**Figure 17: A probability range updated in response to a test result**

The green curve guides the updating in the positive outcome; the red curve would be used to figure the new range had the result been negative. In this example, the prior probability (“likely”, 0.67 – 0.90) is replaced by something close to a “very likely”. The posterior probability, given the new evidence, is narrowed down to [0.89, 0.97], close enough to “very likely”, which starts at 0.90. A stronger test would be represented by a green line pushed up further to the left and to the top and would thus have resulted in a new range yet a bit higher and narrower.

The beauty of this logic becomes more fully apparent in the next diagram. This comes closer to qualitative reasoning. Note that the green and red lines have been replaced with crescent-shaped areas. These express the uncertainty that surrounds the strength of the test. The inside boundary of the crescent, close to the dashed diagonal line, represents a weaker, less discriminating test. Its outside boundary, pushed towards the sides of the square, is the line of a stronger test. The crescent, instead of a simple curve, admits that the strength of the test is not fully understood. This kind of modesty suits expert judgment well.

The six-panel diagram demonstrates the mapping of the prior range to the posterior for the six cases of ( Likely, Uncertain, Unlikely ) X (Positive, Negative Test Result). A positive result always triggers an update to the next higher probability range (incl. the “very likely”). Vice versa for a negative result. But, depending on where we start, the posterior range is narrower or wider than the prior. This change in precision depends on the test result. For example, looking at the two panels in the top row for the “likely” prior probability, a positive result leads to the narrower “very likely” category. A negative results leads to the wider “uncertain” category.
Figure 18: Posterior probabilities in response to six cases of priors and test results

<table>
<thead>
<tr>
<th>Pretest Probability</th>
<th>Positive Test Result</th>
<th>Negative Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Likely</strong></td>
<td><img src="image1" alt="Diagram" /></td>
<td><img src="image2" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Uncertain</strong></td>
<td><img src="image3" alt="Diagram" /></td>
<td><img src="image4" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Unlikely</strong></td>
<td><img src="image5" alt="Diagram" /></td>
<td><img src="image6" alt="Diagram" /></td>
</tr>
</tbody>
</table>

Source: op.cit.: 166

After all that, the diagrammed relationships may look trivial. But it seems doubtful that analysts and experts unfamiliar with Bayesian reasoning would understand them as what they really are. These updated beliefs handle two types of uncertainty:
The prior probability that we assign to an event before the test already is vague.

The quality of the test – i.e. of the new evidence – itself is uncertain.

Nevertheless, rules of translation from prior to posterior are possible and make sense.

This is a methodological achievement. One of the rules advanced by Medow and Lucey is that, for events that are very unlikely or very likely, not one, but two tests are necessary to move our beliefs to the adjacent category, “likely”, respectively “unlikely”. “This is because the change in the probabilities is small within these categories. Two concordant results are needed to change out of the very categories” (p. 166).

Where is the humanitarian application? We see plausible applications primarily in the kinds of thought experiments that are part of scenario exercises. At least two judgments are required:

- The participating experts identify a crucial development – a particular course of the crisis, the expansion of humanitarian action, etc. in a defined future period – and assign it a range of probability.
- Next, they identify potential drastic events each of which would be strong enough to knock off the probability from the initially selected range to an adjacent range.

For example, after a period of unbroken calm and with a growing trickle of returnees, experts observing a conflict zone may review the prospects of mass returns. Suppose the experts no longer exclude “that over the next six months, half of the IDPs will return to their pre-war communities”. They give this proposition a cautious “uncertain”, which is a broad probability range all the way from 34 to 66 percent. But they also anticipate – this is the “test” – that “any violent incident claiming more than five lives over the next two months” would cause them to revise their belief downward to “unlikely”. “More than five” and “two months”, of course, are arbitrary. These precise criteria are useful, not because there is any tested theory underscoring them, but because they strengthen monitoring and commit to serious revision.

Critics may object that humanitarians use this type of logic all the time – there is no need for Bayesians to teach them. Yet this visual method of updated probability ranges borrowed from medical testing has two benefits. First, to scenario developers it suggests that such ranges should be of unequal width – wider in the uninformative center of the scale (e.g., 34 – 66 percent), and shorter towards the highly informative extremes (e.g. 0 – 10 percent). As we have seen, the updating results in wider or narrower ranges, depending on the prior range and on the result of the test. The updated ranges agree reasonably well with the five intervals that Medow and Lucey recommend – see above.

Second, the method reminds humanitarian experts that our theories are not very precise (and frequently not very accurate either). The use of probability ranges is a sign of intellectual humility amid endemic uncertainty. By contrast, academic research often achieves greater precision. To revert to the context of IDPs or refugees returning, conflict studies makes amazingly precise predictions:
“For instance, if a large low-income country with no previous conflicts is simulated to have two to three years of conflict over the 2015–18 period, we find that it will have nine more years of conflict over the 2019–40 period than if peace holds up to 2018. Conversely, if a large low-income country that has had major conflict with more than 1,000 battle-related deaths in several of the past ten years succeeds in containing violence to minor conflict over the 2015–18 period, we find that it will experience five fewer years of conflict in the subsequent 20 years than if violence continues unabated” (Hegre, Nygård et al. 2017).

These are expert judgments, using large event databases and experts’ simulation skills. Their form, however, is such that they cannot be fully translated into the kinds of judgments that experts and decision-makers must work out for Area X in the next Y months. Of course, the researchers can calculate a generic hazard rate for the recurrence of conflict for X and Y, plugging in X’s values on their model variables and Y as the desired time horizon. This estimate may be helpful as an additional input for area experts who work with local information and with pragmatic models. Yet the uncertainty of a specific hazard rate is so large that again a well-informed probability range is the best we can do.

**Process tracing and cause-effect testing**

In this section we report a method in which experts provide *point* estimates of probabilities. This seems to contradict the recommendation in the previous section to admit uncertainty by using probability *ranges*. However, as we shall see, the uncertainty is cleverly recaptured by eliciting a second probability.

One of the common methods to structure qualitative data and thinking is known as “process tracing” (Collier 2011). Process tracing focuses on an observable event and traces the causal chain back in time in order to identify possible causes. It is described as the analysis of the trajectory of “change and causation” (ibd., 823) - how did we get from A to B? This process is specifically relevant to interactive expert judgement, such as when experts meet in a workshop. When experts make the connections and underlying assumptions transparent, they expose themselves to challenges by other participants.

We discuss this method of causal inference at first in its original deterministic version, and then in a recent probabilistic variant with a Bayesian turn. We present a format in which the probability estimates by the experts can be gathered, aggregated and aligned with the typology of process tracing tests.

**Events, links and tests**

The process tracing exercise begins by identifying and classifying variables of interest. These denote events or developments that have a plausible place in the casual chain. What caused the outbreak of civil unrest? Which developments could have resulted in severe and unexpected flooding?

The method is not dissimilar to building a criminal case, with analysts looking for clues and evidence to drive the story. Crucially, analysts are interested in developments and events that have taken place as well as in those that have not been reported, but plausibly should have happened. To illustrate, in a flood disaster, the casual chain may include excessive precipitation (observed) and tampering with the nearby dam infrastructure.
(suspected, but not proven). Collier, in his leading article on the topic, identifies four tests that can be used to classify the importance of the clues along the way (see table; Mahoney 2012: discusses them further).

Table 17: Process tracing test for causal inference

We apply these tests to an investigation triggered by a humanitarian crisis.

Cholera in Haiti – Origins of the epidemic

Some ten months after the earthquake that devastated Haiti in 2010, cholera broke out. By the end of 2016, the epidemic had claimed over 9,000 lives and had sickened some 800,000 Haitians (UNCA 2016). In recent history, no outbreaks had been reported in this country; the origin of this one posed a scientific puzzle. Some scientists supported a “climatic hypothesis”, arguing that a string of the virus had been dormant in coastal waters and that disruptions in the wake the 2010 earthquake caused it to spread. A “human hypothesis” was also advanced, at first by international news organizations. Reporters noticed serious sanitary problems in a military camp of the Haiti UN peacekeeping mission (Frerichs, Keim et al. 2012). The ensuing hypothesis that “UN peacekeeping troops had brought cholera to Haiti” was specific, and also politically charged. A consensus emerged that a better understanding of the origin of the epidemic was “essential for future prevention of similar outbreaks and [to] acknowledge the right of Haitians to understand the events that led to their cholera devastation” (op.cit.).

Hypothesis tests

The hypothesis “UN peacekeeping troops brought cholera to Haiti” could be causally proven in these Process Tracing tests (adapted from Frerichs et al., op.cit., as well as Cravioto, Lanata et al. 2011):

- **Straw in the wind test**: The outbreak started close to the camp housing peacekeeping troops (*the event is of interest but can, in itself, not confirm or reject the hypothesis*)
Aggregation and synthesis - Expert judgment and Bayesian reasoning

- **Hoop test** (“jumping through a hoop”): A peace-keeping contingent from a country where cholera is endemic, Nepal, arrived before the outbreak. *(the occurrence of the event is a requirement for the hypothesis; if it does not take place the hypothesis is rejected)*

- **Smoking gun test**: Pipes from the camp housing the Nepali contingent leaked fecal waste into the nearby river. This river was the source of local cooking and drinking water in the area where the outbreak occurred. *(The hypothesis is confirmed by the observation. However, the hypothesis would not be rejected if this event were not observed)*

- **Doubly decisive test**: Studies of the strain of the disease showed that cholera was likely to have been brought to Haiti from a single source. The strains isolated in Haiti and Nepal during 2010 were a perfect match. *(If the evidence is confirmed the hypothesis is confirmed; else, it is rejected)*.

This type of exercise encourages experts to take a hard look at how compelling the hypothesized causal chain really is: Have we missed key events? How can the events be ranked in importance, with some being essential to the ultimate outcome, and others merely auxiliary? How certain are the links; how sure are we that A had an effect on B?

**From historical reconstruction to anticipation**

Process tracing as a method is not limited to causal reconstruction after the fact. The identification, description and classification of clues by experts lies at the heart of scenario building and even of early warning systems. The basis for this type of anticipatory analysis is further reinforced by Bengtsson and Ruonavaara (2017), who coined the term “contribution analysis”. It “requires the creation of a ‘causal chain’ where each link represents an intermediate outcome, associated with risks that might prevent it from taking place and assumptions that need to hold if the intermediate outcome is to materialise” *(ibid, 2)*. Conflict studies and humanitarian scenario building are fertile fields for contribution analysis; the wide body of research into peace builders and peace spoilers comes to mind *(in the tradition of Stedman 1997)*, which we cannot further discuss here for space reasons.

**Process tracing and Bayesian updating**

The tests administered to the origin-of-cholera-in-Haiti hypothesis produced results with credible high certainty. In many constellations, however, the links between causes and effects will remain uncertain. Additional evidence can reduce the uncertainty, but will rarely eliminate it. Befani and Stedman-Bryce (2016) have advanced a probabilistic version of process tracing, based on Bayesian updating. Their practical applications are in the field of program evaluation, but the method easily transfers to humanitarian expert judgment. Every piece of the evidence considered important for or against the hypothesis – for example: “The outbreak started close to the camp housing peace-keeping troops” - faces two questions:

- What is the probability to find the piece of evidence if the hypothesis is true?
- What is the probability to find the piece of evidence if the hypothesis is false?

Figure 19: Process tracing tests and probable evidence

Source: Befani and Stedman-Bryce, op.cit., 5
The first kind of probability is analogous to the sensitivity of a test, as we saw in the previous section, the probability of a true positive. The second is the obverse of the specificity of the test, in other words, the probability of a false negative. The authors measure the strength of a piece of evidence by the ratio of those two probabilities\(^{28}\).

Depending on the estimates that the experts make of these probabilities, the pieces of evidence can then be assigned to inform one or the other of the four process tracing tests. We propose that these initial probability estimates should be tabulated, as in this fictitious example (see Table 18), and then aggregated, either by the experts’ own consensus or by the analyst. The table should indicate also the type of test that the aggregated probabilities inform. Tests should never be considered absolutely certain; thus all estimated probabilities should be > 0 and < 1, no matter how close they are believed to be to the extremes.

\(^{28}\) Some readers may, at this juncture, expect these questions to run in the opposite direction: “What is the probability that the hypothesis is true [respectively false] if we find this piece of evidence?” This is not the intent here. We challenge the experts to quantify the degree to which they believe that the particular evidence will be found if the hypothesis is true [respectively false]. Statistically minded readers will note that we are dealing with likelihoods. The likelihood is the probability to observe particular data values given the value of a parameter (such as the truth value of a proposition) (Wikipedia 2017g). Figure 19 defines a “likelihood ratio”, as Sensitivity / (1 – Specificity) (and in this figure it is simply: \(y/x\)). This notion of likelihood ratios goes back to diagnostic testing in medicine; readers wishing to familiarize with its logic may consult another Wikipedia article, which has a worked example (Wikipedia 2017h).

However, in the interest of simplicity, we do not adopt the statistical likelihood terminology in this note.
We leave it to the reader to sketch a diagram of the same kind mapping the expert probability judgments exemplified in this table.

To note some finer points:

- **The argument, not the conclusion, dictates the questions:** The formulation of the items in this table is less specific than in the initial discussion, which followed a logic of certainty. In the probabilistic perspective, we find it necessary to say that “the outbreak started close to a camp housing peace-keeping troops” – any peace-keeping troops, not only those from Nepal. Similarly, the cholera strains in the Haiti epidemic matched those isolated in *some country* of origin of the troops”, not necessarily those in Nepal. Therefore, the experts assign higher probabilities to both cases – the hypothesis is true, respectively, false – than they would had the items been specific of Nepal. Looking at any camp and any country of origin weakens the tests, but strengthens the argument.

- **Experts who disagree may each support a different test:** Experts D and E, specialists in genetic testing, disagree considerably about the strength of this piece of evidence. It comes to inform the doubly decisive test only after they find a consensus; the judgments of expert E alone support the hypothesis in a smoking gun test rather than the doubly decisive one (her probability ratio is close to 10, not 20). The judgments of individual experts, therefore, may inform
different tests, or none. The aggregated judgments may align with a test different from those that the judgments of one or of several of the experts create.

- **The elicitation format must anticipate the aggregation method:** As we emphasized earlier, regarding other aggregation methods, if the analyst anticipates that the research question calls for a process tracing method, he must ensure that the elicitation format lend itself to the requisite types of information. The experts must understand what is required of them, and for what purpose and eventual aggregation method. In particular, they must understand that the questions run the way of “If hypothesis X is true, what is the probability to find evidence piece Y?”, and not the other way around.

In sum

Befani et al., op.cit., have rendered great service to expert judgment methodology. They have found a way to extend the logic of testing the contributions of causes to the studied effect in a probabilistic perspective. Most arguments, including those made or assisted by experts, are composed of propositions some of which are uncertain. Evaluating these jointly and in a unified framework makes for stronger conclusions. There is no reason why humanitarian experts cannot adopt the method in their domain.

We have two reservations:

- The authors claim to have found a “*quali-quantitative approach to establish the validity of contribution claims*” (op.cit.:1). Some readers may believe that this is a breakthrough in mixed qualitative-quantitative methods. This impression is natural; for Befani et al. follow a trail that Humphreys and Jacobs (2015) blazed in “Mixing Methods: A Bayesian Approach”. This line of work essentially remains a probabilistic extension of propositional logic. It is a far cry from the world of humanitarian information in which analysts struggle with the messy co-existence of numeric and textual data. In this milieu, analysts (and visiting experts) are sometimes exhorted to wield the magic of some mixed methodology in hopes to bring order and insight, and researchers make ritualized claims to such methods. Compared to this common view of mixed methodology, process tracing with Bayesian updating is a largely quantitative approach. Whatever these authors understand by “qualitative data” is not entirely clear. They seem to mean categorical and ordinal data (ibd., 654); such data conform to easy-to-manage rectangular variables-by-cases tables. This understanding differs from that in the ordinary language of humanitarian decision-makers, analysts and experts, which includes textual information. There are interesting developments in cross-fertilizing mixed methods, particularly in political science (Glynn and Ichino 2015), but these are unlikely to percolate down to humanitarian milieus any time soon.

- The conduct of the process tracing argument resembles legal proceedings in which the analyst is the judge presiding over a court that hears the testimony of experts. “*just like in a judicial trial where evidence is produced in favor or against a defendant*” (Befani et al., op.cit., 6). It is doubtful that the culture of humanitarian agencies will appreciate this style. Humanitarians do assign probabilities to their hypotheses. But rare are those who give pieces of evidence two probabilities side by side for the event that the hypothesis is true,
respectively false. Humanitarians have been using process tracing widely, avant la lettre – the interwoven itineraries of millions of migrants literally force it upon both popular and agency cultures -, but the form of expert judgments may need to stay close to familiar styles of argument for quite a while to come. The challenges of communicating probabilities arise not only on the elicitation side such as with key informants, but also with consumers of scenarios and analyses who read certainty into statements clearly marked as uncertain.

[Sidebar:] Belief networks: Migration patterns in the Sahel
In the preceding pages, we traced the process that led backwards from the cholera epidemic to four events:

- The outbreak started close to the camp of a peace-keeping contingent.
- From the camp fecal matter leaked into the river, the source of cooking and drinking water for neighboring communities.
- Before the outbreak, soldiers from Nepal, a country with endemic cholera, had arrived at the camp.
- The cholera strains found in Haiti and in Nepal matched.

The four form a causal chain with a high degree of belief that they are sufficient to explain the origin of the outbreak:

Nepali soldiers carry Vibrio cholera →
They bring germs to their camp in Haiti →
Pipes leak germs from the camp into the river →
The first Haitian patients imbibe them with water from the river.

There is only this one chain. Had it been interrupted anywhere – e.g., if all the camp’s fecal matter had been safely stored in hermetic underground tanks -, the epidemic would not have occurred.

However, in many causal investigations, we cannot help assuming multiple causal chains, some of which meet, then diverge again en route. Our beliefs in the relative causal contributions thus become difficult to order, let alone to estimate with any given evidence.

A method known as “Bayesian belief network” helps to unravel the complexity (Wikipedia 2017b). The Wikipedia article works the reader through the simplest example with two causal chains that is mandatory for all students of belief networks: Why is the grass wet? The sprinkler is on, or it rains, or both. But – on this hinges the surprise result! – the operation of the sprinkler, too depends on the weather. Hence the conditional probability tables are almost self-explanatory:
Figure 20: A Bayesian belief network

![Bayesian belief network diagram]


For example, the probability that the grass is wet (wet: T = true) when the sprinkler is on (sprinkler: T = true), and it is not raining (rain: F = false), is 0.8.

A non-trivial question, given these probabilities, then is: What is the probability that it is raining when the grass is wet? We leave it to the reader to follow the Wikipedia article to the result, which is 0.3577 or 35.77 percent. What the reader can easily comprehend from the above figure is that the farther we move down on longer causal chains the more complicated will the probability tables get, with 1, 2, 4, 8 etc. rows, depending on the sum of links from the preceding events.

What is also obvious is that questions of interest can be asked in both directions, and that belief networks are excellent tools in the hands of scenario builders. What percentage of the time is the grass wet under current assumptions? How would this change if it rained not 20 percent of all time, but 40 – with all other things remaining the same?

We have not found convenient Excel-based applications for belief network design and analysis\(^{29}\). Therefore, we do not want to go further into the mathematics of this method. Instead, we briefly report on a study that investigated migration flows in the Sahel with fairly complex belief networks and multiple scenarios with simulated outcomes. The humanitarian interest, given the ongoing Mediterranean-centered refugee crisis, is obvious.

In "Using Bayesian belief networks to analyse social-ecological conditions for migration in the Sahel", Drees and Liehr (2015) combine survey data on individuals’ socio-economic situations and migration motives with ecological records. Two samples, each with 300+ individuals, are from a region in Senegal and another in Mali. With slight differences between the two, the authors design likely causal chains from global conditions (extent of climate change and of political and economic development) through environmental and socio-economic conditions to migration motives and hence

\(^{29}\) A list of belief network software can be found at [http://www.wikiwand.com/en/Bayesian_network](http://www.wikiwand.com/en/Bayesian_network).
Aggregation and synthesis - Expert judgment and Bayesian reasoning

to predicted migration patterns. The figure reproduces one of the networks (Mali) with resulting predicted migration patterns (D), given the observed (A, B, C) and artificially varied (E) preconditions. More on these below.

Figure 21: Belief network about migration patterns in an area in Mali

The authors develop four scenarios. Scenario D is the current baseline; the country is considered to be economically stagnant and institutionally unstable, yet with low impact from climate change as yet.

Figure 22: Definition of four scenarios

Thus, under the baseline scenario the conditional probabilities for any variable $X$ in blocks A and B to which an arrow points from one of the factors in block E is simply the observed frequency of $X$. All variables were discretized, for example in “high” vs “low”; thus:

$$p(X = \text{high} \mid \text{Scenario} = D) = \frac{n_{X=\text{high}}}{N}$$

where $N$ is the sample size, $n$ is the count of respondents with $X = \text{high}$, and observed data are survey data or data from environmental measurements such as rainfall records.
The key question is: Where do the conditional probabilities come from in the hypothetical scenarios A, B and C? The answer is: The research team of which the authors are part generated them as expert judgments. We illustrate this with the values for one of the causal links from block E to block A, marked by the fat blue arrow.

Figure 23: Example of a conditional probability table, Mali sample

Subsequently, the study team fed the probability tables to a belief network application and had it compute the predicted migration patterns under each scenario, separately for Senegal and Mali.

One of the main conclusions from their simulations is:

“For both regions the scenarios indicate the linkage between economic conditions and duration of migration. While non-permanent migrations are currently quite common in these areas (accounting for 84.9% of all migrations in Bandiagara [Mali] and 76.9% in Linguère [Senegal]), an improved economic and political situation (scenarios ‘Limitation’ and ‘Prosperity’) is likely to lead to more permanent migrations. Because seasonal and temporary migrations are often linked to economic motives, other motives – especially ‘education’ – simultaneously gain importance. Furthermore, destinations within the region of origin become more likely at the expense of distant destinations” (ibid.: 335).

In other words, if the situation does not improve, the tendency to migrate over long distances will remain at current observed levels, which, for the Malian region, involves between a quarter and a third of the population (see figure 23).

The authors report the robustness of their estimations in detail. One assumes that the study team had little difficulty agreeing on the particular causal relationships – the set of arrows that connect influencing and influenced variables. The situation is much more challenging when experts cannot agree on the basic causal pattern. Bradley et al. (2014) illustrate the challenge with three experts that have different cause-effect beliefs in a humanitarian situation:

“Predicting famine. An aid agency wishes to do some advance planning for its famine relief operations and consults several experts in order to determine the risk of famine in a particular region. All agree that the relevant variables are R: rainfall, Y: crop yields, P: political conflict, and of course F: famine. But they disagree both on the causal relations between the four variables and on the probabilities of the various values that these variables may take. All consider rainfall to be the main determinant of crop yield. However, while Expert 1 thinks that poor crop yield and disruptive political conflict are the main causes of famine, Expert 2 thinks that the causal influence of political
- Expert judgment and Bayesian reasoning

conflict on famine is indirect, via the effect of the disruption of agricultural production on crop yields. Expert 3 considers the relationship between political conflict and famine to be more complicated still, with political conflict both causing famine directly, by disrupting food distribution, and indirectly, through the influence on crop yields. These three opinions are represented in Figure 1 by a set of DAGs. [directed acyclic graphs].

Figure 24: Experts with differing causal assumptions

Bradley et al. conclude that aggregation has to proceed in two steps, at first in qualitative manner, and only then quantitatively. This sounds pragmatic and constructive, but the authors demonstrate the significant conceptual difficulties. This treatment is far beyond the scope of this study. In the real life of agencies, one assumes, the manifest dissension among experts would likely prompt analysts or even decisions makers to step in and impose a causal model that they can defend vis-à-vis their stakeholders.

Figure 1: Expert causal judgments
Humanitarian applications
5. Humanitarian applications

5.1. Introduction - The Analysis Spectrum
ACAPS has sought to help humanitarian analysts structure the process of information analysis. It has borrowed ideas from the intelligence community (Hibbs-Pherson and Pherson 2012) that portray the analysis process as a two-dimensional progression. In one dimension, the analyst advances from data-driven to concept-driven mental activity. In the other, his outlook changes from reactive to proactive. In this progression, the leading questions change as well. One starts out with a basic curiosity about what the information at hand appears to be about. Several rungs further up on the ladder, the understanding is solid enough in order to switch to a prescriptive mode, suggesting not merely what is the case, but what could and should be done (in response to humanitarian needs). The dominant temporal orientation passes from hindsight to insight to foresight – by degrees, if not by principle, because perceptions at every stage involve anticipatory expectations to find distinctions and classifications recurring in the future. There will be day and night tomorrow; the monsoon season will repeat itself next year even if the rains are uncertain; the potential users of the analysis understand the humanitarian lingo today and will understand it tomorrow. Temporal orientations are always multiple.

Figure 25 depicts the sequence of leading questions. Each of them stands for a level of analysis. A warning is inserted – at some point, plausibly at the explanatory analysis level – individual analysis work becomes less productive. The analyst needs to reach out, to other analysts and to experts. The decision to collaborate is, of course, fraught with various risks, notably the loss of time and consistency. But so is analytic isolation.
In this chapter, we use the sequence of analysis level – exploration, description, explanation, interpretation, anticipation, prescription – as a simple ordering device to arrange case studies. We strive to fill every level with one or more meaningful illustrations. We assign a case to the level at which we believe the experts’ work made the strongest contribution. We do not imply that their judgments – let alone the analysis work that experts perform prior to judgment – are restricted to a single level. In fact, experts are at their best when they help analysts (and, when needed, decision-makers) climb safely up and down the Spectrum ladder. Terms of reference and the ethical distinction between predicting and prescribing place the necessary restraints so that experts will not usurp the Spectrum.

5.2. Exploratory analysis: Using experts to find information

When using experts is better, faster and cheaper

The ACAPS “Refugees/Migrants in Europe” project ran from December 2015 to March 2016. Its objective was to provide analysis of the humanitarian priorities in the five countries most affected in south-eastern Europe. The large geographic scope, fluid situation and limited availability of useful secondary data demanded a specific approach to identifying what type of information was available. ACAPS therefore complemented its review of the existing data with the perspectives that experts working on the crisis throughout Europe were offering.
The initial plan was to conduct regular semi-structured interviews with at least five experts per affected country. ACAPS analysts recruited experts who were able to contribute by virtue of their connections with the affected population or response agencies. Individuals working for governments, INGOs, national organizations, national Red Cross chapters as well as for the United Nations High Commissioner for Refugees (UNHCR) were specifically solicited. At the outset, the analysts established credentials and baseline perspectives through individual face-to-face in-depth interviews. According to the methodology, the analysts were to re-contact the experts every 2-3 weeks by phone or e-mail (as preferred by the individual experts); every expert would be requested to update ACAPS on the changing situation. To facilitate comparability over time, the analysts were to ask the same set of questions during every round. However, this level of participation was not reached in practice (see “Application and Challenges”).

Processing and analysis of the data took place on an on-going basis. The analysts would enter the responses into Google Forms, a Web-based instrument that made information storage, sharing and question development easy. During regular project meetings, the analysts would discuss the relevance and validity of the expert data and of other information sources. They would challenge the findings as well as identify key issues and information gaps. This in turn informed the project’s analysis products, including four situation updates, one thematic report and two scenario building reports. Each report was downloaded between 200 and 800 times. In total, four analysts worked full-time for three months to gather, analyze and report on the situation.

Application and challenges

Useful as a light approach to exploring, gathering and verifying information: By the end of the project, 61 different experts from 44 organisations had provided their perspectives. The set-up proved a cost-effective, quick and useful way to identify emerging issues early on. Widespread problems with the set-up of the ‘hotspots’ (EU-run centres that process and host refugees) and overcrowding of accommodation facilities are examples of developments which the analysts picked up before they were widely reported.

Another strength of the approach was that it allowed experts to share views which they themselves would not feel comfortable to put in writing. Examples of such sensitive topics include the appropriateness of the response of different governments/EU institutions as well as bureaucratic constraints hampering the humanitarian operations.

It also allowed for quick comparison of the situation among countries – for instance, experts in FYRO Macedonia indicated early on that most of the basic needs were covered, while those in Croatia warned of significant gaps in services.

The overlap in the roles of experts, decision-makers and stakeholders was considerable. Several of the individuals consulted were both subject-matter experts and members of the target audience of the project’s information products. Strategic decision-makers provided their expert opinions on the situation within their particular countries of operations, while at the same time consulting ACAPS material in order to stay abreast of regional developments. This regular interaction with the main audience proved useful to confirm the appropriateness of ACAPS work and to locate critical information gaps.
However, weaknesses in the design and roll-out of the project constrained the network’s potential:

**Design not adapted to levels of participation:** The project fell short of its ambition to repeatedly interview the same experts. The analysts managed to interview fewer than half of the experts more than once; only five participants responded three or four times. Follow-up was hampered by key informant turnover, lack of incentives and a shortage of analysts, given the intended scope and scale of the project. This limited the information needed to closely follow the evolution of the crisis. In a bid to keep updates easy, ACAPS settled for a questionnaire that turned out to be too light for the task. In hindsight, a more extensive questionnaire, administered to completely changing samples of experts, might have been more productive.

In the social science literature, this question is far from being settled: Vul (2010) and Vul et al. (2014) are among the few investigating sampling of cognitions, and by implication, of experts. Rasli (2006:56) reports that in a multi-round expert opinion assessment in the IT sector, the attrition among experts was significant. Of the 25 experts participating in the first round, 21 continued to the second while only 13 could be retained for the third. The second round failed to produce significant concordance of opinions; the third succeeded (probably because fewer experts needed to work out consensus). There is a small, thematically heterogeneous literature mentioning “rotating experts”, but the meaning of the term varies. Some note that rotating experts are no valid substitutes for standing expert committees.

**Lack of a structured approach to the qualitative analysis:** The main type of quantitative information of interest to this project - the number of people arriving to and residing in the different countries, was gathered from two specific on-line sources. The network of experts was therefore only used to gather perspectives of a more qualitative nature, chiefly concerning humanitarian priorities. There was no set methodology in place to process this unstructured data. The four analysts therefore made decisions on how to synthesize conflicting perspectives, the level of reliability and how to weigh the various expert opinions on a case-by-case basis. This was possible

---

30 Icard (2015) summarizes interesting findings by Vul and Pashler (2008): “More recently, [they] performed a similar study, specifically on point estimation problems like the Africa/UN example […]. in order to test whether averaging multiple guesses by the same subject would be more accurate than the average accuracy of their guesses, as has been demonstrated for groups of subjects (first, incidentally, by Galton 1889). They found this to be the case. Furthermore, averages were more accurate for subjects tested three weeks apart than twice on the same day, suggesting that samples become more independent as time passes. Thus, not only do subjects exhibit random variation, but responses appear to originate from some underlying distribution, which itself may encode more accurate knowledge of the world than any single sample drawn from it. In some sense this latter point is obvious. When making an estimate, say, about the percentage of African countries in the UN, we bring to bear all sorts of knowledge about Africa, the UN, and any other relevant topic, which cannot be captured by a single numerical estimate (for that matter, even if the point estimate were “optimal”). What is interesting about Vul and Pashler’s results is that repeated elicitation of estimates gives evidence that subjects’ intuitive theories of many of these domains are surprisingly accurate, and that these intuitive theories are organized in such a way as to produce samples from a sensible distribution, as the Sampling Hypothesis proposes”.

A voice from the health care sector, Drost (2006) opposes successions of experts in psychiatry: “A complicated succession of experts may visit the patient, asking the same questions and using the same files, distracting his or her attention and motivation from the treatment process. This system is threatened by the qualitative and quantitative scarcity of properly qualified experts”.

118
because of the analysts’ country specific expertise and triangulation of the key informant perspectives with other information sources (albeit limited by the scarcity of secondary data).

One of the main pitfalls of this approach is that it does not mitigate the analysts’ biases. The input of the experts with whom the analysts had developed a relationship was valued more than information from other sources. For this and multiple other reasons - transparency, efficiency and accuracy – similar exercises in future will need a more tightly structured approach.

**Lessons learned**

- Expert judgment can substitute for the dearth of good, available data, both primary (survey) and secondary. ACAPS was able to gather information from experts quickly and cheaply.
- At the outset of multi-round collaborations, the analysts must carefully establish credentials, baseline perspectives, and viable emotional rapport. A satisfactory rapport will motivate experts to verbally share information that they would not commit to writing.
- Consistency over time of persons – on both sides -, motivations and meanings can be a bedevilling challenge. The strategy must anticipate expert attrition, a mechanism for finding quick replacements, and briefing the new experts. One must accept that across rounds of interviews the samples of experts will not be matched.
- Similarly, one must anticipate that there will be analysis challenges. The organization must ensure consistency across its participating analysts and adaptations followed by all of them.

**5.3. Descriptive analysis: The population inside Syria, 2015**

The descriptive analysis addresses questions of the Who, What, Where, When and How Many. Experts make two kinds of contributions to it. Formally, they bring expertise that guides the collection and analysis of observations – in other words, not judgments, but measurements such as through administering survey questionnaires. Substantively, experts fill gaps in data with judgments that they derive from prior knowledge. In somewhat stilted statistical lingo, one might say that substantive experts fill in parameters of multi-variate distributions when the sample of interest is inadequate.

Sometimes experts apply formal and substantive skills simultaneously, such as when they access outside datasets and determine whether their statistics can enrich the analysis of the current data of interest. For example, the sample at hand may provide a plausible estimate of the population mean, but may manifestly underestimate the variance. The expert obtains a credible estimate from some place else that makes for a better description of the population.

**Case: The population inside Syria in summer 2015**

The Whole of Syria Needs Identification Framework (WoSA-NIF) was a collective effort to obtain sub-district level planning figures of population, IDPs and Returned IDPs for the 2016 program year. During spring and summer 2015, a large part of the
humanitarian community, coordinated by OCHA, lent hands. ACAPS supported the effort in the design phase of one of the components – the elicitation of population estimates from key informants inside Syria – as well as during the analysis and aggregation of sets of estimates that originated from three distinct sources.

The WoSA approach was remarkable for the openness with which the participants cooperated with specific expert judgment methods. The limitations of these in this particular context became manifest in the implementation. Specifically, the WoSA planners agreed to try out a probabilistic method that would allow them to gauge the uncertainty around sub-district population estimates as well as around the ensuing aggregate national estimate. The following sections describe initial results, necessary adjustments and, in retrospect, the quality of the information so collected. We limit the discussion to estimates of the population believed to be present in the sub-districts in August 2015; the ways of estimating IDPs and returnees are not the subject here.

**Probabilistic framework**

The WoSA planners placed great confidence in key informants, with the vast majority of whom they would foreseeably communicate only by means of mobile telephony and limited Internet services. This set practical limits on intensity and feedback during training and testing phases. Still, by and large, key informants understood the basic intent of providing their best estimates as well as a measure of their felt uncertainty. Together with the most plausible estimates of their local sub-district populations, they returned their subjective minima and maxima. They thus supplied the data needed for simulating point and interval estimates from the kind of triangular probability distributions that we discussed in the previous chapter (pages 68–74). Ultimately, WoSA partners supplied 1,323 usable estimates on 217 of the 270 sub-districts. The number of estimates per sub-district varied from one to twelve, with a mean of 6.1. The word “estimate” in this arrangement has a special meaning: Each key informant supplied, not one number, but - implied in minimum, mode (most plausible value) and maximum - a probability distribution of the true value of the population in her sub-district.

While that effort was coordinated out of Amman, the OCHA office in Damascus led a series of independent estimates as part of the creation of humanitarian “Governorate Profiles”. This group too elicited minimum, best estimate and maximum for the population of each sub-district. It applied a Delphi method that settled on one distribution per sub-district. It supplied sets of the three triangular parameters for 267 sub-districts.

**Multiple sources for best coverage ...**

In addition, OCHA purchased commercial population estimates based on satellite photography evaluation. The Landscan data (Jordan, Watkins et al. 2012) came without uncertainty measures and would not have qualified but for their higher coverage (269 sub-districts). They were not suitable for the same statistical treatment as the first two sources and forced the adoption of a different aggregation strategy.

In a first step, the WoSA key informant estimates were unified, via simulation, into one probability distribution per covered sub-district. The Governorate Profiles did not need that step because they provided only one distribution per sub-district to begin with. In
a further simulation step, the means from these distributions were estimated, separately for WoSA and Profiles\(^{31}\).

... necessitate multi-method aggregation

The WoSA and Governorate Profile best estimates of sub-district populations were then placed in a common table with the Landscan-based estimates. The data fusion method known as “Beroggi-Wallace” (described in the previous chapter) was used to produce combined estimates. All 270 sub-districts were covered; for 208, estimates were available from all three sources. Two sources were available for the estimates in each of the other 62.

Looking at the sets of best estimates side by side, a committee of experts, the Regional Assessment Team, reviewed them, sub-district by sub-district. In some cases, the Team would revise the estimates, particularly when it had more recent information on sub-districts with besieged communities and those with massive displacements.

Substantial upward bias

It became apparent that the combined estimates produced a national aggregate that far exceeded the upper limit of the national population that could be deduced from credible country-level statistics. The Beroggi-Wallace method yielded a national estimate of 19.3 million people. A simple demographic model, based on consensus figures, produced a probable figure of 16.4 million.

This latter figure was obtained as the difference between the population in 2013 and the outflow of refugees from official Economic and Social Commission for Western Asia and UNHCR statistics (ESCWA 2015, UNHCR 2015). Other forms of migration and excess mortality were not taken into account. Neither were the natural population growth and returnee refugees considered because it was assumed that their sum was far lower than the sum of migrants (most of them unregistered refugees outside Syria) and excess deaths. Therefore, this estimate defined an upper bound on the current population; it was highly unlikely that the true population figure should exceed it.

Table 19: Estimate of the national population in 2015

<table>
<thead>
<tr>
<th>Source</th>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESCWA</td>
<td>Population 2013</td>
<td>18,180,000</td>
</tr>
<tr>
<td>UNHCR</td>
<td>Refugee stock 31 Dec 2013</td>
<td>2,301,668</td>
</tr>
<tr>
<td>UNHCR</td>
<td>Refugee stock 25 Aug 2015</td>
<td>4,089,023</td>
</tr>
<tr>
<td></td>
<td>Outflow</td>
<td>1,787,355</td>
</tr>
<tr>
<td>Calculated</td>
<td>Estimated population 2015</td>
<td><strong>16,392,645</strong></td>
</tr>
</tbody>
</table>


Two consequences ensued. An uncertainty measure was needed for this global national population estimate. Second, the sub-district estimates needed to be adjusted, to correct for the manifest upward bias.

\(^{31}\) Why the means were estimated by means on simulation, and not simply calculated as (min + mode + max) / 3, see also the explanations in the previous chapter.
Uncertainty of the national estimate

Around the alternative estimate of 16.4 million people, a 96-percent confidence interval was wrapped. Its width was in proportion to the relative distance of lower and upper bounds from the means in the simulated WoSA national estimate\(^3\). The same exercise was done also with the Government Profiles-based estimate.

To exemplify with the WoSA estimate, which was the highest among the three sources:

The mean of the simulated total population in the 217 sub-districts that the WoSA key informants covered was 23,441,402. The lower bound of the confidence interval was 21,183,203, which was equal to \(0.9037 \times 23,441,402\). Thus 0.9037 was the multiplier to obtain the lower interval bound for the alternative estimate of 16.4 million. The WoSA upper bound was 26,123,188. This gave the multiplier for the upper bound as \(26,123,188 / 23,441,402 = 1.1144\). These multipliers were then applied to the alternative estimate. The same was done with the factors computed from the Governorate Profile estimate. This table gives the alternative estimate and the two variants of its confidence interval.

Table 20: Confidence intervals around the population estimate

<table>
<thead>
<tr>
<th>Estimated population 2015</th>
<th>16,392,645</th>
</tr>
</thead>
<tbody>
<tr>
<td>relative to:</td>
<td>2pcLB</td>
</tr>
<tr>
<td>Gov. Profile simulation</td>
<td>15,996,074</td>
</tr>
<tr>
<td>WoSA simulation</td>
<td>14,813,483</td>
</tr>
</tbody>
</table>

LB: lower bound of the confidence interval, UB: upper bound. 96 percent confidence.

Not surprisingly, the confidence interval from the Government Profiles is narrower, reflecting the Delphi consensus in the constituent sub-district estimates. The WoSA’s is wider; the uncertainty among its multiple key informants per sub-district was mathematically passed on, rather than behaviorally reduced. However, the Government Profiles working group did not document the steps in its Delphi rounds. One is thus inclined, in an abundance of caution, to prefer WoSA’s, based on fully reproducible operations.

Sub-district estimates: Adjusted and revised

All sub-district population estimates were adjusted uniformly, and subsequently some were revised individually.

Given the unacceptable upward bias in the national aggregate under the Beroggi-Wallace method, the initial sub-district population estimates could not be used as absolute figures. Rather they were now seen as structural estimates, assumed to stand for proportions in the national population. In this sense, they were the skeleton for adjusted estimates. To illustrate how the adjustments were made: The initial estimate of the Damascus sub-district, combining the WoSA, Governorate Profile and Landscan estimates, was 1,728,127 persons, part of the estimated national population of

\(^3\) The 96-percent choice, as opposed to the conventional 95 percent, was the consequence of an arcane technicality in the statistical application used in the simulations. It is of no importance here.
Humanitarian applications - Descriptive analysis: The population inside Syria, 2015

19,279,782. This national estimate, as we have seen, was replaced by the more plausible 16,392,645. The proportionate adjustment for Damascus thus worked out as 1,728,127 * (16,392,645 / 19,279,782) = 1,469,341. Analogously for the other sub-districts. Around these adjusted estimates, confidence intervals were wrapped, with bounds obtained by the multipliers for WoSA and Governorate Profiles.

Two objections to this re-allocation procedure leap to the eye. First, the adjustment factor (16,392,645 / 19,279,782) ≈ 0.85 was constant; that implied that the extent of the upward bias was the same everywhere. This was unrealistic. The construction of confidence interval bounds using constant multipliers on the point estimates was a similarly unrealistic uniformity.

There were simply no instruments in the collected data to make the adjustments more flexible. This in itself already justified the review by the Regional Assessment Team, whose members, from multiple perspectives, added relevant new and specific information. For 31 out of the 270 sub-districts, the estimates revised by the Team were outside the confidence intervals based on the WoSA multipliers. For 33, the revised estimates were outside the Governorate Profile-derived intervals. This may mean one or both of two things: Most of the estimates, once they were adjusted so as to sum to the 16.4 million national figure, seemed plausible in the eyes of the Team members. Just as likely, the detailed review of 270 sub-district figures exceeded the Team’s capacity; instead, they focused on the minority that were debatable or had changed in known significant degrees.

Lessons learned

Different groups of persons contribute as "experts"

The standard assumption of a homogeneous group of individuals all working on the same footing did not apply in this case. Neither may it work in other settings with complex information processes.

By the time the ACAPS analyst combined estimates from three sources, contributions from several sets of experts had been mingled. The designers of the Needs Identification Framework formed one such set, familiar with agency formats, information requirements and, specifically through ACAPS, the opportunity to capture the uncertainty of the estimate right from the elicitation point.

The Damascus-based group detailing the Governorate Profiles were a related and, in part, overlapping set. A host of partner agencies relied on their field staff inside the country as key informants; these persons were experts in the own right and place. The experts that produced the Landscan population estimates lived in their own universe, never met by any of the other groups. It was yet another type of expert, a Geographic Information System (GIS) specialist in OCHA Amman, who translated the Landscan data to a useful format.

The decision-makers, if we want to consider the leaders of the Whole of Syria (WoSA) approach and their seniors as such, also acted as experts. They vouched for the official validity of the key figures that went into the demographic equation that produced a plausible upper bound on Syria’s population in 2015.
Thus, while classic expert judgment theory posits experts from homogenous, if differently weighted backgrounds, in real life experts move as a motley crowd, in a parade with shifting members.

**Individually, key informants are too sure; collectively, they are cautious**

Adjusting the estimates from the WoSA key informants was necessary because they resulted in a figure for the national population that was too high. It by far exceeded the upper bound calculated from UNHCR and ESCWA statistics. Some key informants may have exaggerated their estimates in hopes that higher numbers would result in higher relief allocations. This seems plausible, but motives are impossible to prove on that basis.

Apart from the upward bias, one thing that we can statistically demonstrate is that the individual key informants were too sure in their beliefs. Typically, the maximum value offered was only 20 percent higher than the minimum. In many sub-districts with turbulent population movements, such narrow margins underplayed the extent of uncertainty to which even well informed residents were subject. Not surprisingly, then, within a given sub-district with multiple key informants, the extremes of the aggregated opinions were much farther apart – the 98-percentiles were typically (= median ratio) 2.2 times the 2-percentiles.

Paradoxically, in this large uncertainty there is value. The value is in having, thanks to the diversity of key informant knowledge, *reasonable measures of uncertainty*. The high uncertainty among WoSA key informants was more credible than the comparatively narrow ranges that the Government Profiles working group produced. And, what is more, the uncertainty measures derived from key informant estimates could be aggregated upwards. In this case, estimates were aggregated to governorate and national levels, for which realistic confidence intervals were calculated from the simulated aggregate distribution. The estimated national population of 16.4 million was bracketed with the confidence interval [14.8 million, 18.3 million], which implied a precision of roughly ±10 percent. Given all the unknowns, this seems to be a reasonable conservative stance.

The value of the multiple key informant opinion collection is thus not in the absolute numbers of the raw estimates. It is in realistic measures of confidence gained in the analysis.

**The need to improvise**

On the drawing board, elicitation designs and analysis plans for expert judgments may look neat and final. Imperfections or outright failures arise in the execution. If the collected information still serves part of the objectives, experts and analysts may need to improvise.

This happened also in the sub-district population estimates. The example is of an arcane technical nature, but it illustrates how a seemingly minor deviation painfully slows down a process that must be completed against a deadline.

It was in the advanced stages of programming the algorithm to aggregate multiple WoSA key informant estimates that it became clear that some estimates referred only to part of the sub-district area. This happened chiefly in sub-districts over which two or
more belligerent parties held control. Some key informants ventured estimates only for their own “side” of the area. This created an entity issue, with the basic entity “sub-district X key informant” unraveling to “sub-district X area (complete or part) X key-informant”. The fix in the simulation code was laborious and, in hindsight, less than optimal. Also, while re-programming, there were multiple updates communicated on the extent of area covered by some informants, requiring update merges in the statistical application. Accommodating another level of aggregation – from part to the whole of sub-districts – and repeated updates roughly doubled the programming time.

Elicitation and analysis are not written in steel and marble. They are buoyed by the turbulence of the larger – cognitive and social – task environments in which they take place. It is due to the strength, creativity, and tolerance for improvisation, of these environments that eventually we arrive at whatever useful results.

5.4. Explanatory analysis: Problem tree Ebola

Widening the perspective

Ebola broke out in Guinea in December 2013, was internationally recognized as an epidemic ravaging also Liberia and Sierra Leone in spring of the following year and was declared closed in May 2016. By this time, it had claimed more than 11,000 lives (Wikipedia 2016g). At the beginning of the crisis, the international community perceived the Ebola outbreak in West Africa as a purely public health emergency. The response was oriented towards the containment of the epidemic and treatment of the sick. The initial focus of the response was on providing beds for patients and mobilising health practitioners. The livelihoods, education or protection needs of the affected communities, indirectly caused by the outbreak, were left unaddressed.

However, it soon became clear that the secondary humanitarian impacts of the epidemic were extensive, threatening the lives and livelihoods of more than 22 million people in the three most affected countries, Guinea, Sierra Leone and Liberia. The disruption of public and private services created an “emergency within the emergency”. Entirely focused on disease control, humanitarian actors failed to activate their surge capacity, or set up emergency funding and coordination structures. It took time for the humanitarian community to recognise the complexity of the crisis and respond to the secondary impacts.

The ACAPS Ebola project

The lack of reporting and understanding regarding these impacts prompted ACAPS to take on a corrective advocacy role. From the outset it was obvious that efforts to address critical needs beyond epidemic control would struggle with enormous complexities. A thorough, comprehensive and widely shared analysis of impacts and needs was in order to better support the affected population and to guide the humanitarian response. ACAPS set up an analysis project looking specifically at the humanitarian implications of the epidemic and involving ACAPS staff in Geneva, Guinea, Liberia and Sierra Leone. They worked in close contact with NGO, UN agencies and government authorities on the ground and at headquarter level to ensure that its analysis would be useful and used by the Ebola responders.
Use of a structured analytical technique

A major lesson learned during this epidemic has been the need to broaden the scope of the humanitarian response during a large-scale Ebola outbreak. In hindsight, this appears common-sensical, if not always obvious. At the time, however, it took concerted efforts to change perceptions, priorities and ultimately resource mobilization. In this collective endeavour, ACAPS’ analytical approach proved helpful.

Just as it was critical for the epidemiologists to isolate the Ebola strains, it was important to unravel the entire variety of secondary impacts and of the derived unmet needs that together with the epidemic formed the humanitarian emergency. However, social causes and effects are not as neatly demarcated as biological organisms. Often, multiple factors are responsible for a given humanitarian condition, but some have more weight than others. Grappling with this slippery complexity, the analysts resorted to a tool known as the “problem tree”.

The graphic device: The problem tree

The problem tree is a graphic structure – formally a directed acyclic graph ("DAG"; see Wikipedia 2017d), in which there are no cycles – in the center of which the analyst places a focal or “core problem”. The core problem is a summary of the mechanisms that link phenomena perceived as “causes” to others seen as “effects”. The causes are layered among themselves, with primary causes activating secondary, tertiary, etc. ones. Similarly, effects are layered, with the immediate ones from the core problem cascading into follow-on effects. The cascades on both sides represent the beliefs of the analyst who conceptually isolates causes and effects. The figure exemplifies a possible qualitative structure. The strength of our beliefs in the individual links can be quantified such as in Belief Propagation Models (Krause and Clark 1993, Wikipedia 2017c); these were neither feasible nor necessary in the advocacy that ACAPS directed to the humanitarian issue in point and are therefore not further discussed here.
Of course, in a given policy arena, there can be multiple core problems, each motivating its problem tree. The integration of the trees can be problematic. All indications in real life agree that problems can be in mutual cause-effect rapport. Thus the effects of Problem A may be part of the causes of Problem B, and vice versa. If so, the graph becomes cyclical, exceeding the canonical problem tree representation. Some cyclical graphs can be reoriented to become acyclic again, through a process called “condensation” (Wikipedia 2017d, op.cit.), but it is unlikely that any combined core problem graph of the secondary Ebola impacts would have conformed. Instead, the ACAPS analysts found a different representation.

**The ACAPS problem tree**

In order to catalogue the kaleidoscope of secondary impacts, the analysts initially moved away from graphs with vertices (boxes) and edges (arrows). They used a matrix structure in which humanitarian sectors were assigned to columns. Five sectors - Health, WASH, Food Security, Livelihood and Protection - were profiled. In a second dimension (the rows of the matrix), slots were created for five categories of problems: Availability, accessibility, awareness, quality and utilization. The intuition that guided the categorization was, first, that deteriorations in any of those aspects would frustrate the fulfillment of essential needs. Second, the aspects were sufficiently distinct as well as exhaustive, in order for identified specific problems to be mapped to one or the other. The categorization is similar to one proposed for health services by the UN Committee on Economic, Social, and Cultural Rights in 2000, which comprises availability, accessibility, acceptability, and quality (UN-CESCR 2000, discussed in Gruskin, Bogecho et al. 2010):
Availability
Availability refers to the physical presence of goods and services in the area of concern through all forms of domestic production, commercial imports and aid. It is determined by the production, trade, stocks and transfer of goods in a given area. Availability problems could include a lack of facilities, staff or goods.

Accessibility
Accessibility refers to people’s ability to access and benefit from goods and services. It often concerns the physical location of services, but can also be influenced by economic or security restrictions. Accessibility problems could include services located far away from people’s homes, lack of transportation, high fees or insecurity.

Awareness
This refers to whether people are aware of the existence of goods and services. If services exist but are not visible or known to residents, then the need may be for an information campaign rather than the creation of new services. Awareness is determined by the message appropriateness, the communication channels and the frequency of updates. Awareness problems could include a community not knowing or having incorrect information about where they can obtain free medication.

Quality
The quality of services depends on the number of people with the required skills and knowledge to perform a given service, the capacity to deliver the service with the fewest resources and in a timely manner, and the feeling of security for beneficiaries. Quality problems could include ineffective or below standard services.

Utilisation
The extent to which goods or services can be used to achieve specified goals with effectiveness, efficiency and satisfaction in a specific context. Utilisation is determined by the knowledge of the individuals on how to use a specific good or service, the attitude and beliefs of a person towards that good or service and the actual practice of these beliefs. Utilisation problems could include the fear of going to the hospital to get treatment.

During summer 2014, four ACAPS analysts built a cumulative catalogue of causes, problems and effects, filling a series of evolving matrix representations. During a workshop in August, the version shown in this figure emerged. For lack of time, it concentrated on problems of availability and accessibility. This was the time when it seemed far from certain that the epidemic could be contained. The problem matrix was construed on the somber background that the disaster might explode to a larger scale, and with it the secondary impacts.

Figure 27: Problem tree of the expected Ebola impacts
(next page)
Humanitarian applications

EXPECTED IMPACT OF THE EBOLA OUTBREAK CRISIS ON HUMANITARIAN SECTORS

**HEALTH**
- **AVAILABILITY**: Inadequate availability and coverage of health services
  - Lack of limited health personnel
  - Lack of limited medical equipment
  - Lack of limited medical supplies
  - Lack of limited IPE and protective materials for health staff and health centers
  - Lack of limited health infrastructure
  - Lack of limited lab capacity
  - Lack of limited access on emergency services
  - Non-ebola dead body management
- **ACCESSIBILITY**: Physical and logistic constraints (fuel, transport, etc.)
  - Security, safety constraints (tensions, quarantine, etc.)
  - Lack of limited access to money, credit, money, financial services
  - Fear
  - Other: use of traditional healers, patients living in the bush, prison population

**WASH**
- **AVAILABILITY**: Lack of limited water and sanitation services available (fuel, electricity to operate)
  - Lack of limited water trucking in some areas
  - Lack of limited waste management collection or services
  - Lack of limited chemicals for water treatment
  - Lack of limited domestic and personal hygiene items (chlorine, soap, etc.)
  - Lack of limited number of body bags
  - Lack of limited protective equipment for sanitation and body management teams
  - (disinfection materials for corpses, personal protection, etc.)
  - Lack of limited sanitation team human resources
- **ACCESSIBILITY**: Physical and logistic constraints (fuel, transport, movement restrictions, etc.)
  - Security, safety constraints (tensions, quarantine, confinement, movement restrictions, etc.)
  - Lack of limited access to money, credit, money, financial services
  - Fear

**FOOD SECURITY**
- **AVAILABILITY**: Lack of limited food production
  - Lack of limited food stocks at household level
  - Lack of limited food stocks available in markets
  - Lack of limited fuel for cooking
  - Lack of limited meat - increased use of wild bush meat
  - Lack of limited income
  - Lack of limited access to credit
  - Seasonality and conflict
  - Deterioration of livelihoods

**LIVELIHOODS**
- **AVAILABILITY**: Lack of limited agricultural livelihoods
  - Lack of limited livestock inputs (livestock, animal feed, vaccination, etc.)
  - Lack of limited formal and informal traders stocks (food, agricultural produce, livestock inputs, small business and informal traders stocks, etc.)
  - Lack of limited demand for services (Transport, business, tourism, etc.)
  - Lack of limited international trading volume (minerals, agro products, etc)
  - Lack of limited tourists and international visitors
- **ACCESSIBILITY**: Physical and logistic constraints (fuel, transport, movement restrictions, etc.)
  - Security, safety constraints (tensions, quarantine, confinement, movement restrictions, etc.)
  - Lack of limited access to money, credit, money, financial services
  - Fear

**PROTECTION**
- **AVAILABILITY**: Lack of limited law and order resources
  - Lack of limited access to public services
  - Lack of limited protection services and safe spaces
  - Lack of suspension of classes
  - Lack of limited personal protection equipment
  - Lack of limited civil registration and legal services
- **ACCESSIBILITY**: Physical and logistic constraints (fuel, transport, movement restrictions, etc.)
  - Security, safety constraints (tensions, quarantine, confinement, movement restrictions, etc.)
  - Lack of limited access to money, credit, money, financial services
  - Fear

**DETERIORATION OF THE HEALTH STATUS OF THE POPULATION**
- Increased morbidity and mortality

**DETERIORATION IN NUTRITIONAL STATUS OF THE POPULATION**

**DETERIORATION OF LIVELIHOODS AND INCOME**

**DETERIORATION OF PROTECTION STATUS OF THE POPULATION**

**REDUCED PUBLIC AND PRIVATE SERVICES**
- **DISPLACEMENT**: Forced or voluntary, internal or across borders
- **REDUCED MOBILITY**: Forced or voluntary
- **POLICY AND GOVERNMENT MEASURES**: Border restrictions, quarantine, curfew, forced declaration of ebola cases, etc.
- **SOCIAL, CULTURAL PRACTICES AND BELIEVES**: Lack of trust in institutions, traditional healers, dead bodies management and funeral practices, media (radio, etc.)
The problem tree as advocacy tool

This problem tree served as an advocacy document to approach donors and finance a specific project dedicated to the analysis of the Ebola crisis. It was also used as a basis for documents produced subsequently. The proposed priority sectors to consider were identified through this matrix. Thanks to the preliminary identification of potential problems linked to the large-scale Ebola outbreak, ACAPS staff were able to accelerate the production of reports on health, food security, WASH and protection. They were able to quickly analyse and produce reports on these secondary consequences.

Further work

This problem tree was not updated during the height of the crisis. However, towards the end of the crisis, in January 2016, the ACAPS analysts reviewed the observed problems and impacts of the Ebola outbreak on the different humanitarian sectors. The following month, ACAPS published a 39-page lessons-learned report “Beyond a public health emergency: Potential secondary humanitarian impacts of a large-scale Ebola outbreak” (ACAPS 2016a). It detailed, with numerous observations and sources, likely disruptions of services and access to goods, to provide a non-exhaustive plan to help future responders. The report condensed the insights that the analysts had gathered in 70 interviews with members of the United Nations, NGOs, donor organisations, national responders and academics as well as from a literature review. Past instances of disruptions were ordered in three steps – sector, problem category and specific problem. This table renders a small segment from the “Health sector – Availability” category.

Table 21: Segment of a problem tree on secondary impacts of a large-scale Ebola outbreak

<table>
<thead>
<tr>
<th>Problem category: Availability</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Funding</strong></td>
<td>Routine health service expenditure suffered during the Ebola outbreak. In September 2014, only half of the planned amount was received. Cost recovery dropped due to the decrease in use of health services and some donor funding for the health sector was redirected to contain the epidemic (UNDP, 23/12/2014).</td>
</tr>
<tr>
<td>Decrease of non-Ebola health service expenditure</td>
<td>Some hospitals were entirely taken over by Ebola patients or had to close because they could not treat Ebola patients safely (international media, 25/09/2014). In 2014, clinical teams and facilities for tuberculosis were repurposed to the Ebola response (WHO, 10/12/2014). In October 2014, in Liberia, Doctors Without Borders (MSF) had to close its 200-bed referral hospital near Bo because of the strain of responding to the Ebola outbreak (MSF, 16/10/2014).</td>
</tr>
</tbody>
</table>
Three things are remarkable about this report:

1. It modified the generic problem tree format in the scheme of sectors, problems (corresponding to the five problem categories), and impacts (specific problems, some further detailed in cause-effect propositions).
2. It determined that the secondary impacts of a large-scale outbreak would manifest themselves in five priority sectors: health; water, sanitation and hygiene (WASH); people’s safety (aka protection); education; food and livelihoods. Education was newly added; food and livelihoods were combined.
3. It established plausibility for every specific problem foreseen, with one or several documented historic instances.

Figure 28: Modified problem tree, example demand on health services

Source: ACAPS (2016a:6)
Lessons learned

Uses and usefulness

Problem trees should be regarded as a “thinking instrument, rather than a final product” (Veselý 2008). They are useful in structuring the analysis and experts’ thinking, rather than by providing a “scientific position on a problem”. They are particularly useful at the initial stage of the analysis when they help experts to dissect a broad issue into smaller units, one causal “root” and “branch” (effect) at a time. The problem tree can particularly help in:

- distinguishing between different parts of a problem and grouping them
- identifying the nature of elements and the relations among them – is a specific issue a cause, a focal problem or an effect?

In hindsight, the Ebola problem tree could have been more widely used and disseminated to alert the humanitarian community of the potential effects of a large-scale Ebola outbreak. Instead of being solely a fundraising tool, it could have been used to raise awareness and advocate for more multi-sector assessments on the impact of the crisis.

Steps to design a problem tree

- Step 1: Formulate the core problems. A problem is not the absence of a solution but an existing negative state: ‘Crops are infested with pests’ is a problem; ‘No pesticides are available’ is not.
- Step 2: Select one focal problem and place it at the centre of the tree.
- Step 3: Arrange causes and effects in a hierarchy. Causes will be placed below the core problem while effects will be above.
  - Identifying and mapping immediate and direct causes of the focal problem.
  - Identifying and mapping immediate and direct effects of the focal problem.
- Step 4: Identify the secondary causes, the causes of the immediate causes. Do similarly for the effects. You will likely move causes around a lot to be able to determine if they are primary or secondary causes.
- Step 5: Review the problem tree to verify its validity and completeness.

While the problem tree can be done by one expert, it is preferably done as a group exercise, to help gather different perspectives. This may be limited by workshop time and ability to find a consensus.

A willingness to adapt formats is required. The original template, as a network with boxes and arrows, may not always work, depending on the type and volume of material to represent. For the Ebola problem tree, a matrix format worked better. It accommodated an categorization of problems (availability, accessibility, awareness, quality, utilization) already established in public policy.
5.5. **Interpretative analysis: Severity ratings in complex emergencies**

**A severity scale for northern Syria**

Rapid assessments after disasters gauge the intensity of unmet needs across various spheres of life, commonly referred to as "sectors". Several assessments have used measures of needs proposed by ACAPS. Commonly, two measures are taken - a "severity score" independently given to each sector, and a "priority score", a relative measure comparing the level of needs in a given sector to those in other sectors. Needs in every assessed locality are thus scored twice.

In the first half of 2013, ACAPS was involved in three rapid needs assessments in Syria. These were the Joint Rapid Assessment of Northern Syria, assessing the situation in 58 of the 128 sub-districts in six northern governorates of Syria (ACU 2013), the Joint Assessment of the city of Aleppo, covering 52 urban neighborhoods (AWG 2013a), and the second Joint Rapid Assessment of Northern Syria (J-RANS II), extending the area assessed to 106 sub-districts (AWG 2013b).

The Second Joint Rapid Assessment of Northern Syria (J-RANS II)\(^{33}\), published in May 2013, employed severity scales in five sectors (public health, food security, nutrition, shelter, water and sanitation). Key informants in sub-districts were asked to express their beliefs on a five-level severity scale. They were asked which of the following statements best described "the general status" of the sector:

<table>
<thead>
<tr>
<th>Table 22: The severity scale used in the J-RANS II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No concern – situation under control</td>
</tr>
<tr>
<td>2. Situation of concern that requires monitoring</td>
</tr>
<tr>
<td>3. Many people are suffering because of insufficient [supply of goods or services]</td>
</tr>
<tr>
<td>4. Many people will die because [supply of goods or services] are insufficient</td>
</tr>
<tr>
<td>5. Many people are known to be dying due to insufficient [supply of goods or services]</td>
</tr>
</tbody>
</table>

**Combining expertise for assessment quality**

An important facet of this assessment was that each enumerator was assigned one sub-district to assess, based on her experience and knowledge of the area. In most sub-districts, enumerators spent enough time (5 to 12 days, depending on the size of the sub-district) in order to interview several groups of key informants. They used the same type of questionnaire in every meeting and ultimately created their personal synthesis for the sub-district in yet another copy.

\(^{33}\) Available at [http://reliefweb.int/sites/reliefweb.int/files/resources/JRANS%20II%20-%20Final%20Report_0.pdf](http://reliefweb.int/sites/reliefweb.int/files/resources/JRANS%20II%20-%20Final%20Report_0.pdf)
Finally, the enumerator brought back her results to the assessment cell, where trained debriefers would review her report and give special attention to the severity ratings and the priority ranking.

Of interest to our expert judgment study is the fact that the final severity ratings for each sector and visited sub-district reflected discussions, validations and agreements across several layers of expertise:

- The expertise of multiple key informants in the visited sub-districts. They provided insights into the current situation of sectors, areas and in population groups of which they were the most knowledgeable.
- The expertise of the enumerator, who progressively built situational awareness as her interviews and visits progressed. She also made direct observations and applied her recollections of previous times to comparing pre-crisis conditions to current ones.
- The expertise of the two debriefers, who interviewed the returning enumerators and reviewed the assessment questionnaire. They would ask questions about each and every severity score for each sector and each location. They would not accept synthesis questionnaires at face value; the enumerators had to actively argue for the credibility of their findings. A briefing would commonly last more than three hours, with the enumerator explaining on a map where she had been, with whom she discussed, what she did see and didn’t see, defending her synthesis point by point, even to the point of presenting additional evidence in the form of photos if security had permitted to take any. The debriefers would home in on the plausibility of the severity ratings, adding personal knowledge of the area and results from other debriefings. It happened several times that by the end of the session debriefers and enumerator would agree to modify a severity rating.
- The expertise of the analyst, comparing and mapping results from all assessed areas, finding outliers and inconsistencies, and asking for specific crosschecks where it was necessary.

**Lessons learned**

Valuable lessons were learnt from the J-RANS II process and fed into later assessments in Syria, i.e. the Syria Integrated Needs Assessment (SINA), Multi-Sector Needs Assessment (MSNA) 2013 (AWG 2013c) and MSNA 2014 (Humanitarian Liaison Group 2014).

**The severity scale:** Despite a thorough and thoroughly documented validation process, the severity scale of the J-RANS II produced data of limited value. The designers had not anticipated that in areas already heavily devastated by the conflict the five-point scale had little discriminatory power. Subsequently, ACAPS proposed a version of the severity scale with seven levels (Benini 2013:48-49):

Table 23: A seven-level severity rating scale

When you consider the situation in the xxx sector, would you say

1. There are no shortages
2. A few people are facing shortages
3. Many people are facing shortages
4. Shortages are affecting everyone, but they are not life-threatening
5. As a result of shortages, we will soon see some people die
6. As a result of shortages, some people have already died
7. As a result of shortages, many people have already died.

Evaluative criteria in the analysis process: Assessing data quality and drawing analytical conclusions are distinct processes; they call for different tools and methods. Evaluations take place at three levels: of each piece of evidence, of the body of evidence, of the inferential processes. Different criteria apply, particularly between the first two and the third level. For the benefit of the Syria assessment teams, ACAPS devised this diagram of levels and evaluative criteria:

Figure 29: Changing evaluative criteria, by stages of the analytic process

Mixing expertise: J-RANS II fortuitously brought together national and international experts, the former knowledgeable of Syria and its contexts, the latter trained in humanitarian analysis. The mixture was effective. So was the organization of multiple analysis layers, fostering the progressive elucidation and contextualization of assessment findings. The controlled process mitigated biases and elevated the discussion from the quality of the data to the quality of the conclusions.
5.6. **Interpretative analysis: Clarifying priorities**

**Coordination in the early stages of the disaster response**

On April 25, 2015, an earthquake struck Nepal, killing about 9,000 people. Within days, the United Nations Disaster Assessment and Coordination (UNDAC) activated the Nepal Assessment Unit in Kathmandu\(^{35}\). The Unit was in charge of collating secondary data and providing daily situational analysis to various humanitarian stakeholders active during the earthquake response. It was composed of assessment specialists from OCHA and ACAPS. Also, it benefitted from secondment of staff from agencies already present in Nepal before the crisis, such as the World Bank, the United Nations Development Program (UNDP) and Microsoft Lab, who volunteered their time for the assessment unit.

**Collating and reviewing information**

During the first month of the crisis, the cardinal event in the Unit’s life was the updated briefing package, released every morning at 11h00. The package included an analysis of available national and international news and social media messages as well as specialized humanitarian reports. With vast amounts of new information accruing with every day, the package provided a synthesis of the most current knowledge of humanitarian conditions\(^{36}\).

Every member of the Unit was assigned a specific source stream to monitor on a daily basis. The reviews of new information had to be ready by 09h00, in time to select key elements and trends for the consolidated briefs at 11h00. One analyst would survey international news media, another followed the national media, yet others were bent on cluster documents and minutes, on social media (Facebook, Twitter, etc.), and on national authorities. The graph below details the information flow from sources to products during the first month of the crisis.

---

\(^{35}\) UNDAC is part of the UN emergency response system for sudden-onset emergencies. It is designed to help the United Nations and governments of disaster-affected countries during the first phase of a sudden-onset emergency and assists in the coordination of incoming international relief. For more information see: [http://www.unocha.org/what-we-do/coordination-tools/undac/overview](http://www.unocha.org/what-we-do/coordination-tools/undac/overview).

At its peak, the Assessment Unit employed 17 people. Collating and structuring information took up most of the capacity. Only two or three persons did full-time analysis work and report writing.

Specialization by information sources was a response to the strong time pressure. This organizational form had its own advantages and downsides:

**Advantages:**
- Over time, the members learned to identify the most interesting and relevant sources and to discard less pertinent ones. Thus they found that only a small number of international newspapers were providing relevant content to humanitarian decision-makers. By focusing on specific media streams, the analysts could recognize significant reporting patterns.
- Continuous individual work with the same sources made it easier to spot duplicate or amplified stories. For instance, stories of one case of suspected cholera were reported six times and subsequently re-appeared as a story of six cases.
- Systematic and broad media coverage would allow members to identify a large spectrum of situations and differentiate between rare and extreme situations or conditions, generally over reported by international and national media for mediatic effects and more common conditions, less frequently reported.

**Disadvantages:**
- The arrangement fostered segmented knowledge. Every assessment unit’s member had a detailed and up-to-date view of what he/she could glean from their sources, but found it hard to see the big picture. It took organized efforts to build broader views, not unlike a jigsaw puzzle that makes sense once fully assembled.
By extension, the analysts chiefly worked with the stories that their assigned sources carried, as opposed to issues that were reported elsewhere or not at all. Important information gaps may have gone unnoticed.

When similar information arrived from different streams of information (e.g. local news that subsequently was picked up by international media channels), it took time, energy and occasionally good luck to detect the redundancies, determine the most credible version and avoid duplicate records.

Gathering expert perspectives
The “big picture” effort gave rise to another major daily event. Every day at 13h00 the Unit members would gather and share information and knowledge. These meetings followed a set procedure:

1. A session facilitator without sector/cluster mandate was assigned to facilitate each session. Each session would last one hour maximum. The facilitator would make sure to have a good representation of insiders (people with experience of Nepal or past disaster experience in Nepal) and outsiders (people with experience in natural disasters but no specific experience in Nepal). The facilitator would recall the rules before the meeting started. Specifically, members were to refrain from judgment or personal conclusions/assumptions in terms of severity of conditions or priorities. The focus was on getting the facts rights and to check figures, types of evidence and sources supporting new information.

2. Every member would give a summary of the most significant news in her media stream, supported with numbers of articles, and how they would repeat across different media in her assigned media stream.

3. Members would report trends and patterns that they had identified in recent days - stories that were still repeating and others that meanwhile had disappeared. Novelty was valued, particularly about newly accessible geographical areas and affected vulnerable groups or sectors.

4. Others who had similar or related information from their own source streams were encouraged to intervene and complement the analyst’s highlights. They were free to challenge an opinion by asking for more information or details.

5. After the individual updates were heard out, the group would proceed with a ranking of affected areas and of priority sectors. Every analyst would name, in plenary, the three geographical areas that he considered the most impacted in the light of the accumulated information. Also, he would vote for the three sectors that, in his view, should receive assistance with the highest priority. Individual who would not feel comfortable in providing their judgment could pass.

These choices were recorded on a white board. They were not replaced by a consensus vote as in a Delphi process. Rather they were aggregated using a formula known as the “Borda count” (Wikipedia 2011b) and recorded in the Unit’s log.
The evolution of beliefs in the most affected areas is manifest in this graph where color gradient and size of bubbles represent the degree of priority given to a specific district. The degrees were measured by the Borda counts (Wikipedia 2011b) of the priority votes by the Unit members every day from 2nd to 11 May 2015.

During this period, priorities shifted from districts with a wealth of information (Gorkha, Sindhupalchok, Dhading) to districts (Rasuwa, Nuwakot and Lampung) with wide initial information gaps that were progressively closed as more reports reached Kathmandu.

**Lessons learned**

For the elicitation of expert judgment, the Assessment Unit developed a number of practices that are of interest beyond this particular disaster:

- The pragmatic and transparent approach to identifying priority needs and priority districts worked well in this dynamic environment. In fact, it worked well under time pressure and strict production schedules. It worked because the Unit followed a “good enough” philosophy that privileged timeliness and conclusion over the quest for the best and the most complete.

- Resources, knowledge and expertise present in Nepal were combined with specialized (e.g. assessment methodology) assets brought in from outside. This is the organized and encouraged dialogue and discussion between both insiders and outsiders that mitigated potential biases and provided with “the best available version of the truth so far”.

- The Unit set up an effective forum for discussion and exchange around priority needs. Every member had the chance to make her case, on an equal footing with all others, and casting the same number of votes on priority sectors and areas. The daily meeting would increasingly be joined by new participants, eager to hear about the last updates, if not to voice specific concerns or issues from their own professional circles.
Politics neutral, methodically rigorous and transparent to the outside, the Unit shared findings that the users found credible.

Above all, the work of the Unit taught a lesson in humility. Findings were only as good as the information on which they were based. Everybody learned early on that new information could dramatically change perceptions and interpretations, frequently in a matter of one day. A fluid environment makes for fluid judgment.

5.7. **Interpretative analysis: Information-poor environments**

**Mapping information gaps in post-hurricane Haiti**

The Category-4 Hurricane Matthew, which passed through Haiti on 4 October 2016, resulted in the largest humanitarian crisis in Haiti since the 2010 earthquake. It had a devastating impact on infrastructure and people’s homes and livelihoods. About 2.1 million people were affected, of which 1.4 million were in need of humanitarian assistance. According to the preliminary results of the Post Disaster Needs Assessment, the overall damage and losses have been estimated at US $2.8 billion. The hurricane’s impact added to pre-existing humanitarian needs throughout the country, notably related to the cholera epidemic, the El-Niño-induced drought, the migration crisis and the displaced following the 2010 earthquake (Humanitarian Response Plan January 2017-2018 01/2017).

On October 5, 2016, the United Nations Disaster Assessment and Coordination (UNDAC) requested ACAPS’s analysis support. Two ACAPS analysts were deployed to Port-au-Prince for four weeks as part of the UNDAC assessment team to analyze the humanitarian needs.

However, the political climate in the country was tense; the national authorities were determined to prevent the recurrence of shortfalls in government oversight and in coordination with humanitarian agencies that had tarnished the earthquake response in 2010. As a result, humanitarian information was no longer freely shared; it first needed to be validated by the government before publication. The policy made it hard to obtain detailed disaggregated information; this in turn hampered the prioritization of affected areas, necessary for a timely response. Humanitarian decision-makers struggled with significant information gaps.

**ACAPS information gaps methodology in the context of Haiti**

Earlier in 2016, ACAPS had released an extended note on information gaps in multiple needs assessments in disaster and conflict areas (Information gaps in multiple needs assessments in disaster and conflict areas) (Benini, Chataigner et al. 2016). Information gaps are perceived shortfalls of received information against expected information. In the same note, ACAPS investigated the practicalities of measuring information gaps, proposing a scale to rate the usability of information (pp 38). In order to highlight geographic and sectoral gaps in the three most affected departments (Grand’Anse, Sud and Nippes) in Haiti, ACAPS developed this instrument further. The output was a series of maps lifting out the areas where information was critically short.
A first map was produced on 12 October by simply scoring the availability of information by sections communales (admin level 3) in two departments (Grand’Anse and Sud), on a scale of 0 to 4, 4 being the best score.

**Map 1 – Information gaps at section communales in Grand’Anse and Sud, as of 12 October 2016**

The information available on each commune and sector was evaluated and rated against three criteria - quantification of people in need, analytical value and geographical areas covered. Table 24 describes criteria and rating definitions. Note that higher scores mean higher information value.

**Table 24: Haiti information gap scoring system**

<table>
<thead>
<tr>
<th>Score</th>
<th>Quantification of people in need</th>
<th>Analytical value</th>
<th>Geographical areas covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Sector not covered</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>No people in need estimate</td>
<td>Unmet needs</td>
<td>Only main city</td>
</tr>
<tr>
<td>2</td>
<td>Verbal qualifiers only</td>
<td>Unmet needs and causes</td>
<td>Only easily accessible areas (coastal areas, along main roads...)</td>
</tr>
<tr>
<td>3</td>
<td>Rough people-in-need (PIN) estimates</td>
<td>Unmet needs, root factors and priority issues</td>
<td>Only urban areas</td>
</tr>
<tr>
<td>4</td>
<td>Good PIN estimates</td>
<td>Unmet needs, root factors and intensity</td>
<td>Some urban and rural areas (not representative of the whole commune)</td>
</tr>
<tr>
<td>5</td>
<td>Precise PIN estimates</td>
<td>Unmet needs, root factors, intensity and priority issues</td>
<td>Representative of all communes</td>
</tr>
</tbody>
</table>
The measures of “Quantification” and “Analytical Value” indicate how “usable” the information on that specific area is. They let the analysts see how much information speaks to a particular sector, and what that information can tell:

- **The quantification** score indicates the level of precision. This ranges from No meaningful people-in-need (PIN) estimate (e.g. “large numbers of people are in need of …”), verbal qualifiers only (e.g. “most/thousands of people are in need of…”), rough magnitudes (e.g. “Over 10,000, or between 10 and 15 thousand”), good estimates (e.g. “Around 54,500 people in need of…”) all the way to precise estimates (e.g. “10,161 people were in need of…”).

- **Analytical value** refers to the depth and range of the information provided on a given sector within an assessment. It is based on the ability to access the core components of information: Unmet needs, root factors contributing to needs, prioritization of needs and intensity of needs.

The ratings were constructed as follows. A team of experts reviewed all secondary data available. The team of experts then rated, on the three scales described above, every combination of relevant topic (humanitarian access and sectors) and affected commune (admin level 2). An intermediate total score per topic was then calculated, taking the median of the three indicator values in the given commune. Finally, the total score of a commune (“Total Median”) was calculated as the median of the intermediate scores. No weighting was applied to any sector or commune.

By 17 October 2016, thirteen days after the hurricane hit Haiti, the team had given 61 communes * 9 sectors and special topics * 3 criteria = 1,647 scores (1,646 were recorded). Here is a segment of the information gaps database, with the total score in the second column:
### Results

At that point (17 October), of the 61 assessed communities, 11 earned a total median of zero; 33 communities scored 1; 16 scored 2; and one scored 4. On none of the communities was there enough precise information available to earn it a score of 4.

Medians, by definition, drive summaries away from extremes. It is informative to consider the distribution of all the 1,646 scores at the elemental level. Zero was given 565 times (34%), “1” 477 times (29%), “2” 323 times (20%), “3” 252 times (15%), and “4” 29 times (1.8%). The relative frequencies differed by criterion:

- **For Quantification**, 73% of the scores were zero or one, meaning that for three quarters of the communes, two weeks after the hurricane, no PIN estimates were available.
- **For Analytical Value**, 83% of the scores were equal or below Level 2. That meant that for the majority of the communes, intensity and priority questions could not be answered yet.
- Finally, for **Areas Covered**, the same percentage of scores were equal or below Level 2, meaning that for the majority of the communes, only easily accessible areas were assessed while information about the other urban and rural areas had not yet reached the humanitarian agencies.

An update was produced a week later, on October 23. From 13 communes with no information earlier at all, only six remained in this condition. The other seven graduated...
to level 1 or higher. By this time, the information on one commune qualified for level 3, while all the others remained at lower levels, similarly to the week before.

[Sidebar:] Is there a pattern in the commune information profiles?

Our summary of the information levels that the assessment attempts had achieved, by October 17, gives the impression that they were similarly poor across all topics and communes. This was not the case. Particularly the education and, in smaller measure, the shelter/NFI sectors attained higher levels more often. It is therefore of interest to know how the 61 communes differed in terms of their information gap profiles, and what the drivers were for the major differences.

The statistical method chosen, given the ordinal nature of the scores, is the so-called Q-factor analysis (Donner 2001, Raadgever, Mostert et al. 2008). Our adherence is partial, in the sense that the scores were generated without the normalizing format that is mandatory in Q-sorts preceding the factor analysis.

In our case, Q-factor analysis turns the 61 communes into variables (they become column heads), and the $9 \times 3 = 27$ topic X criterion combinations into observations (ultimately only 26 matter, because of a missing value that disqualifies “Shelter and NFI X Areas covered”). We exclude the access-related combinations because we suspect that they follow an entirely different dynamic (this is disputable).

The overwhelming impression is that differences in commune information profiles are primarily driven by the opposition of education scores vs. the scores of shelter/NFI and displacement. The commune-level correlations between items of the opposing groups are mostly negative, e.g.

### Table 26: Haiti - Cross-tabulation of the analytical value of information, two sectors

<table>
<thead>
<tr>
<th>Displacement Analytical Value</th>
<th>Education Analytical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

Total: 61

\[ \text{gamma} = -0.61 \quad \text{ASE} = 0.178 \]

Individually, the correlations are not all of them strong. But the scores on the first statistical Q-factor are clearly opposed (in this Q-factor analysis the items have scores, the communes have loadings).

Why assessments that produced higher information value on education tended to produce, in the same communes, lower levels on displacement and on shelter/NFI, is not understood, absent any knowledge of institutional dynamics. It is plausible that in areas of higher physical destruction, assessment teams looked chiefly into shelter and displacement, and had little interest in collecting data on impacts on education. And vice versa.
Is this the only driving force that differentiated commune information profiles? Not necessarily, but it was by far the strongest. When statisticians extract distinct factors, they look at a statistic known as “eigenvalues” (Wikipedia 2017f) and generally give serious considerations to all factors with eigenvalues greater than one. With these data, the eigenvalues (before rotation) were (rounded): 28, 11, 7, 5, 3, 2. The difference between the first and the second eigenvalues is so large that the game is practically over with the first factor. It accounts for 62 percent of the variance.

As ever so often, statistics helps to spot interesting differences, but it cannot explain them without external knowledge.

Information gaps and decision-making

The ACAPS team analyzed the information gaps during the prioritization of areas for subsequent food distributions. It is acknowledged that the team’s report maps prompted the humanitarian community to ramp up assessments although ACAPS did not learn the details of how particular organizations responded. The map of October 17 was widely referenced in publications, was noted in government meetings, and was perceived as useful by the humanitarian community.

The process of scoring information gaps

Even though some guidance was provided, the scores were given subjectively; they did not leap from any mathematical formula, such as “if more than two entries are about health in commune A, then the score is 2”. More entries do not equal better data. Rather it was up to the two experts to assess the strength of this information, in a subjective manner.

- **Step 1**: A secondary data review on Haiti had previously been done by the ACAPS experts. Both had been reading, selecting key information, tagging it by sectors, severity and geographical area and storing it in an Excel spreadsheet for several weeks.
- **Step 2**: Both experts filtered the information stored and tagged per commune and then per topic to see how much and what type of information was available.
- **Step 3**: Each expert rated the information available on one criterion for one specific topic and commune.
- **Step 4**: Both experts discussed their scores and presented their arguments justifying their opinion.
- **Step 5**: Together, both experts then attempted to agree on the score for each criterion. When they differed, e.g., when one expert opted for level 4 while the
other preferred level 1, they would spend extra time searching for new information. Then, they would present new arguments backing up their preferences. In this small group of two, few disagreements occurred; usually they were resolved by “splitting the difference”.

- **Step 6:** The two experts repeated this process for each sector for each commune.

As the scoring system relied heavily on the experts’ opinions, background and experience were key. Both experts had researched, reviewed and tagged all the secondary data available and were therefore quite intimately familiar with the information landscape. Both had arrived in Haiti shortly after the hurricane hit; both had in the meantime produced situation reports. Both were familiar with the ACAPS information gaps methodology and knew the criteria used in the past.

**Lessons learned**

- Extensive knowledge of the available data is necessary in order to guide the expert judgment analysis.
- Separating the problem of information gaps in smaller units with an indicator attached to it helps the experts coming up more easily with a score.
- Working in pairs or small groups helps in deciding scores by exchanging perspectives and confronting opinions.
- Previous knowledge of methodologies on assessing information gaps is necessary to understand the criteria and making educated calls.

5.8. **Anticipatory analysis: Qualitative risk analysis**

**ACAPS Risk Analysis**

The objective of ACAPS Risk Analysis is to augment the scope of forward-looking analysis and contribute towards the analytical quality of humanitarian early warning and preparedness. It is intended to assist decision-makers in planning response measures based on the perceived likelihood of the risk occurring and its potential impact on humanitarian needs. ACAPS Risk Analysis focuses on contexts that are deteriorating beyond their current trend. Risks identified are expected to materialize within a one to six-month timeframe.

ACAPS Risk Analysis culminates in the risk reports. The risk reports are produced on a monthly basis and contain more detailed analysis of specific risks that are considered particularly relevant or dynamic. For both the risk reports, ACAPS outlines the rationale for each risk, an estimation of likelihood and impact and a breakdown of the potential impact across each humanitarian sector.

ACAPS Risk Analysis differs from other case studies in this document in that it relies primarily on secondary data, rather than heavily relying on the opinions of external experts. A team of ten analysts that works part time on this project, reviews the secondary data, and in this case are responsible for providing the expert judgment.
Application and challenges
The process of risk analysis is supported by, and feeds into the daily work schedule of, the analysis team as each analyst already follows between 6 – 10 countries over a sustained period of time. The process therefore depends on a team of analysts who, in order to provide expert judgment, have an in-depth contextual understanding of a range of countries. Contexts are monitored on a daily basis; past risks are researched; trends are identified and pre-crisis and baseline information is considered. During the process of identifying risks, collaborative thinking is necessary to push ideas and to deepen the level of analysis. Both internal group meetings and external consultations with other agencies facilitate this. Collaboration and discussion with others is essential for challenging ideas, ensuring that all angles have been considered, as well as helping to validate opinions and support the analysts’ judgement.

The seven steps of ACAPS’s risk analysis:

- **Problem analysis**: understanding the context, including pre-crisis information and vulnerabilities as well as understanding of the current situation and drivers of a crisis.
- **Identification of risk factors** present in a certain country, guided by a set of topics to assess vulnerabilities.
- Analysis of the **likelihood** of the identified risks, with the help of indicators adapted to the context.
- Analysis of the potential **humanitarian impact** of the identified risks, based on expected affected areas, historical precedents, and similar situations.
- **Inclusion or exclusion of the risks**, taking into account predefined criteria and thresholds.
- **Prioritization** of the risks presented in the list, based on the combined likelihood and potential impact of each risk.
- Continued **monitoring** of the risk and potential adaptation if the analysis changes.

The risk analysis project has, however, been a work-in-progress since its inception and it remains so, as challenges continue:

- **Harmonization**: It is difficult to ensure a harmonized approach to risk identification, and estimation of likelihood and impact, across a team of ten analysts.
- **Identification and monitoring**: Identifying risks in a timely manner can be difficult due to time pressures and a lack of available information. It is also difficult to establish the difference between a deterioration and progression in the current trend.
- **Measuring likelihood**: Making a judgement call on the likelihood of the risk occurring has proven to be difficult, as has quantitatively assessing the impact and assigning the impact a severity score. Likelihood describes the certainty that a particular outcome will happen, and is therefore a subjective measure.
Likelihood is usually expressed in words such as “likely” or “unlikely”, while probability does the same with numbers, often with values between 0 and 1 or with percentages. As the exact probability of future events occurring cannot be mathematically computed, ACAPS uses likelihood statements. In the Risk Analysis project, these are distributed on a five-point scale as follows.

**Figure 32: The risk analysis likelihood scale**

<table>
<thead>
<tr>
<th>Highly Unlikely</th>
<th>Unlikely</th>
<th>Somewhat Likely</th>
<th>Likely</th>
<th>Highly Likely</th>
</tr>
</thead>
</table>

- **Measuring impact**: The objective of estimating impact is to determine the most likely humanitarian consequences if the risk occurs. The impact should include the magnitude and the intensity of humanitarian needs. The magnitude should be expressed as a number of people requiring additional humanitarian assistance. This includes people who previously did not require humanitarian assistance, people whose existing needs increase in severity.

**Table 27: Risk Analysis impact scale**

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Minimal</th>
<th>Stressed</th>
<th>Serious</th>
<th>Crisis</th>
<th>Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>20,000 - 50,000</td>
<td>Very Low</td>
<td>Very Low</td>
<td>Very Low</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>50,000 - 150,000</td>
<td>Very Low</td>
<td>Very Low</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>150,000 - 300,000</td>
<td>Very Low</td>
<td>Low</td>
<td>Moderate</td>
<td>Significant</td>
<td>Significant</td>
</tr>
<tr>
<td>300,000 - 500,000</td>
<td>Low</td>
<td>Moderate</td>
<td>Significant</td>
<td>Significant</td>
<td>Major</td>
</tr>
<tr>
<td>500,000 and more</td>
<td>Low</td>
<td>Moderate</td>
<td>Significant</td>
<td>Major</td>
<td>Major</td>
</tr>
</tbody>
</table>

**Inclusion or exclusion**: The criteria for inclusion or exclusion of a risk is difficult to consistently define (see table 28 at the end of this case study).

- **Removal**: It is difficult to establish at what point a risk should be removed from the list: for example, it is difficult to determine exactly when the risk has occurred and is therefore by definition no longer a risk. A further grey area is that while a risk may not have occurred within the initially suggested timeframe, it may be deemed just as likely to occur in the coming months. It is difficult to assess whether the risk should be removed or placed elsewhere (see example in the table at the end: Burundi).

- **Lack of data**: As the process relies primarily on secondary data, it can be difficult to judge whether or not a risk remains valid if the circumstance is not being reported on. In some cases, information may only be available when a
risk has already occurred, in which case it can no longer be considered a risk (example: Libya).

- **Short production timeframe**: As the risks are produced regularly, this forces a very quick review. This is problematic when assessing the change in likelihood, as there is a tendency to focus too much on the present moment rather than the broader picture.

- **Confidence**: Aside from methodological difficulties, there are issues concerning analysts’ confidence in their own expert judgement. There is a tendency to err too much on the side of caution out of fear of being wrong. This can lead to overlooking situations that do develop into risks because there is a reluctance to make what may be interpreted as overly bold statements (example: Kashmir).

### Lessons learned

In order to address some of the challenges faced, the need to follow a more sophisticated methodology became apparent. The purpose of following a more concrete methodology for ACAPS Risk Analysis is not to make the process overly technical, but rather to develop a structure that helps to guide thinking and better harmonize the approach towards identifying, monitoring, including and excluding risks as well as defining thresholds to help establish consistency when measuring likelihood and impact. Ensuring a more comprehensive and structured thought process behind risk analysis is also intended to help build confidence in the analysts’ own ability to make a judgement call.

In order to comprehensively understand how a situation evolves from week to week, it became necessary to create a monitoring system that tracks even slight developments within a given context. This is difficult considering time constraints and the limitations of secondary data. However, a list of indicators helps to mitigate some of the above-mentioned challenges. A comprehensive list of indicators can help to remind the analyst of the types of developments to be aware of (e.g., protests or deployment of additional military personnel). This can help the analyst to consider a wide range of factors that may point towards an eventual risk developing. Tracking indicators over time helps to encourage the analyst to take a step back from the immediacy of the situation and consider the broader implications. It also reduces the chance of analysts being overly reactive to sensationalist media stories.

### Table 28: Problems faced in the ACAPS Risk Analysis

<table>
<thead>
<tr>
<th>Example</th>
<th>Challenge</th>
<th>Decision made by analyst</th>
<th>Lesson learned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colombia: Conditions deteriorate for migrants stranded at Colombia – Panama border</td>
<td>The total number of people likely to be affected was low.</td>
<td>Remove – impact considered too low.</td>
<td>Assign thresholds to impact scale (proportional number of people affected): number affected can be based on population figures and historical precedents</td>
</tr>
</tbody>
</table>
5.9. **Anticipatory analysis: Scenario building**

**Leveraging expertise to identify possible developments**

Scenario building is the art of defining the possible ways in which the future may unfold. This is by nature a multi-disciplinary task. It requires expert input, with developments in the political, economic, security and humanitarian spheres all having the potential to influence the future. For instance, in order to understand how a drought might evolve, the input of a meteorologist is required. To be able to then grasp the possible future impact of the drought on households, the collaboration of agricultural specialists, economists, government officials and WASH experts is essential.

To facilitate scenario building, ACAPS applies the “chain of plausibility” approach, which purports to reveal plausible connections across states/events that act both as causes and as effects. A chain of plausibility starts by identifying assumptions that are likely to spark a chain of events resulting in a humanitarian impact. In subsequent

---

**Example**

<table>
<thead>
<tr>
<th>Kashmir: Spike in human rights violations as violence escalates.</th>
<th>Challenge</th>
<th>Decision made by analyst</th>
<th>Lesson learned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of confidence: Analyst doubted own judgment that the situation would escalate. Situation did then escalate.</td>
<td>Omitted</td>
<td>Create a monitoring table and list of indicators to be alert to all warning signs. Provide training to help harmonise approach and make analysts more confident.</td>
<td></td>
</tr>
</tbody>
</table>

| Burundi: Political violence deteriorates along ethnic lines | Risk had not occurred within estimated timeframe but remained likely to still occur. | Moved to long term risk | Create and use list of indicators to help better assess likelihood of when a risk is most likely to occur. Refer back to indicator list regularly throughout the week. |

| Libya: Intensified clashes over the liberation of Sirte | Lack of data available. | Risk was monitored for 10 weeks despite limited data. It was then removed because information about crisis response became available showing that the risk had already materialized. |

---

38 The theory of plausible reasoning is ancient, originating in legal argumentation and mathematical induction. The Wikipedia article is helpful (Wikipedia 2017i), particularly the eleven points that summarize the theory of Collins and Michalski (1989). An explicit theory of plausibility chains is absent. Hints are found mostly with legal scholars. For McCarty (1997), an argument relying on a chain of plausible inferences is weaker than one built on valid inferences that are grounded in authoritative rules. Walker (2006), similarly inspired by legal argumentation patterns, developed a generic diagram...
Humanitarian applications - Anticipatory analysis: Scenario building

steps, the likely impacts in every scenario are elaborated in the greatest feasible detail. Finally, the scenarios are characterized on two simple (five-level) rating scales, for likelihood and impact.  

**Topics and dissemination**

ACAPS has initiated and facilitated several scenario workshops on subjects as different as the impacts of the insurgency in North East Nigeria (2016), of the El Niño/La Niña on the food security situation in Indonesia (2016) as well as of the European Refugee Crisis (2015). Other workshops laid out possible developments of the humanitarian tragedy in Syria (2015). For these exercises, ACAPS analysts elicited and coordinated the contributions of experts in different disciplines; beyond the type of experts usually found in humanitarian action, they occasionally reached out to specialists in domains as exotic as EU border management (European refugee crisis) and the fuel market (Syria). Typically more than 30 experts would be involved in an exercise, mostly through face-to-face meetings.

ACAPS publishes workshop findings in reports (usually about 15 pages long) that describe the most relevant scenarios, their possible triggers and likely impacts. We estimate that ACAPS typically invests approx. 0.75 person-months of staff time per scenario exercise. The scenario reports published so far have been downloaded between 50 and 800 times. The actual number of users is assumed to be higher; the dissemination statistics do not take into account those who received reports by e-mail or in hard copy.

---

Walker (2006:4)

Example of a generic plausibility schema:


ACAPS’ scenario-building exercises hitherto have not used Walker’s formalism. It is doubtful that they are compatible with it. His propositions are strictly true or false; ACAPS’ are probabilistic.  

---

For more information on the scenario building methodology, see the ACAPS Technical Brief on scenario building (August 2016). Probabilities of scenarios are the subject of the next case study.
Lessons learned
While expert judgement is essential to building scenarios, the approach also poses specific challenges:

Different structures for different purposes: Not all of the ACAPS initiatives to gather expert perspectives on the future can be considered a success. Some workshops did not produce any useful material, despite hours of discussion. In others, the scenarios omitted important developments or failed to create buy-in on key components. These, and more productive experiences, have led to the development of a proposed set-up for each scenario-building step, moving from discussions on general topics to the highly specific.

Table 29: Steps of the scenario-building process

<table>
<thead>
<tr>
<th>SCENARIO STEP</th>
<th>QUESTION</th>
<th>EXAMPLE RESPONSE</th>
<th>IDEAL SET-UP</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDENTIFY VARIABLES</td>
<td>General: What are the main current and possible future developments which could impact humanitarian needs?</td>
<td>Level of conflict, rainfall, market prices for staple food</td>
<td>This general question requires input from experts with a generalist view, ideally country/region experts. Bilateral meetings are an effective way to gather such input; bilateral meetings encourage ‘out of the box’ thinking, which group settings tend to inhibit.</td>
</tr>
<tr>
<td>IDENTIFY CAUSE AND EFFECT</td>
<td>Specific: How could these specific developments evolve in the next 6 months? For each scenario, what will be the humanitarian impact?</td>
<td>Example Scenario: The outbreak of conflict in area A, coupled with below average rainfall results in significant decrease in livelihood opportunities, with agricultural production specifically impacted. Displacement, an increase in the prices of basic necessities and the disruption of income sources will increase the number of people in need of support.</td>
<td>These specific questions, which require the combined input of experts with different backgrounds are best addressed in a workshop setting with 10 to 15 experts. A summary of the outcomes of the bilateral meetings is shared with experts, after which they have the opportunity to share their knowledge on specific items (e.g. will rainfall be below or above average) and complement their story with the expertise of others. Invite different experts on the same subject-matter to facilitate cross-checking of perspectives.</td>
</tr>
<tr>
<td>RATE LIKELIHOOD AND IMPACT</td>
<td>Highly specific: What needs to happen before the scenario materializes</td>
<td>It is improbable that an outbreak of armed conflict will occur at the same time as below average rainfall (defined as X standard deviations from the mean)</td>
<td>These specific questions which deal with rating likelihood and impact are best done in a small group of experts (max 5) who have an understanding of both the main issues and principles</td>
</tr>
</tbody>
</table>
Humanitarian applications - Anticipatory analysis: Scenario building

Adapt the type of questions: Technical experts mostly know a lot about a little as opposed to generalists, who know a little about a lot. It is therefore important to deconstruct every problem into very specific questions. For example, during a workshop in Indonesia, it became clear that a question along the lines of ‘imagine there is below average rain in the next three months, what will be the immediate and long-term impact on households eastern Jakarta’ provided more useful responses than asking an agriculture expert ‘how could the situation in Indonesia evolve during the next three months?’.

Disentangle position from perspectives: In some ways it is easier to encourage people to let go of organizational and political biases during scenario building. It is less obvious to represent a certain agenda for a hypothetical, uncertain future. However, this is not always the case, and specifically not if the future is heavily influenced by the conduct of one or more actors participating in the exercise. A scenario exploring the government’s inability to adequately respond to a possible future disaster might prove a sensitive topic if government officials are in the room. The hypothetical nature of scenario building does however provide some space to introduce extreme events to mitigate existing biases. In several workshops the Government’s inability to provide sufficient support was linked to a hypothetical major earthquake affecting the capital. Controversial discussions on the extremely limited capacity of the Government were thereby avoided.

Gather all views: Just as the future, scenario building can go into a multitude of different directions. To illustrate, a scenario building workshop covering northern-Syria in 2015 focused mostly on the price of fuel and the strategy of the Russian government; the participants deemed those factors capable of altering the humanitarian situation within a short period of time and to significant degrees. In hindsight, this specific focus of this discussion meant that the voices of experts in the fields of agriculture, WASH, armed groups, and others were not heard as clearly as was desirable. In order to harness valuable input from all experts, it is productive to assign special tasks to those with special expertise outside the main stream of the discussion. In order to validate the

40 Assigning probability occurs in two steps: first, the probability of the required triggers is identified (e.g. the outbreak of conflict within the next 3 months has a 30% chance, below-average rainfall a 20%). Afterwards, the probability of the scenarios is fine-tuned by comparing the position of the scenarios as relative to other scenarios identified (e.g. scenario A is more likely to occur than scenario B).
specialized experts’ judgements, they should be integrated in plenary discussions during the workshop itself.

5.10. Anticipatory analysis: Visualizing impact and probability

What good are probabilities in scenario-building?
Cooke, a classic in expert judgment, is not favourable to using probabilities in scenario analysis. Borrowing from Kahn and Wiener (1967:6), he defines scenarios as

“But hypothetical sequences of events constructed for the purpose of focusing attention on causal processes and decision-points. They answer two kinds of questions: (1) Precisely how might some hypothetical situation come about, step by step? and (2) What alternatives exist, for each actor, at each step, for preventing, diverting, or facilitating the process?”

A “scenario is salient, not because of its probability but because of its relation to basic trends. If we bear in mind that saliency has nothing to do with prediction, then focusing on a few scenarios might help study the basic trends. However, it is easy to be misled into believing that scenario analysis yields predictions. [Therefore] probability can play at best a subordinate role in the scenario-analytic approach to policy analysis” (Cooke 1991:12 and 11).

So, why have ACAPS-led workshops assigned probabilities to the scenarios that they developed?

We try to give reasons and discuss challenges by reporting on one such workshop. We also look at the visual expression of probabilities and, for that matter, of assumed impacts.

In August 2016, ACAPS, together with the Nigeria International Non-governmental Organization (INGO) Forum, facilitated a scenario-building exercise with humanitarian personnel engaged in the northeastern Nigeria conflict zone. The challenge of assigning probabilities to the scenarios that emerged in that event led to reflections that we wish to share in this sidebar.

Assigning a probability to a future occurrence is fraught with difficulties. One of the premises of scenario-building is that it is not prediction of the future. Scenario-building is most useful when participants are able to consider a range of “futures” without being hindered by arguments as to how implausible they may (or may not) be. Assigning probabilities to scenarios may cause readers to concentrate on only those considered more likely and ignore those less likely.

That would be a mistake as the actual future rarely develops exactly as predicted in any one scenario. Humanitarian response that is informed by as wide a range of scenarios as possible is more likely to be resilient. For example, most humanitarian programming aims to keep costs as low as possible, and usually envisions the programme being undertaken in one particular environment that remains stable (i.e. the security, access,
Humanitarian applications - Anticipatory analysis: Visualizing impact and probability

stability of population, etc. stay the same). If the security situation changes, as it often does, or the target population moves, the programme must adapt. If such changes have not been foreseen during the planning stages, it will be luck whether the programme is able to fulfil its objectives in the changed environment.

Nevertheless, it is useful to give an indication of the likelihood of each scenario as they are not all equally likely. This matters in the context of the cost-benefit and risk analyses that contingency planning necessitates: If the measures identified to make a programme resilient to a certain particular future are costly, they may not be worth undertaking if the risk (impact x probability) associated with that future is low.

**Workshop dynamics**
The elicitation of probabilities and their aggregation in scenario-building workshops pose organizational and facilitation challenges. There is a constant trade-off between the extensive time needed to discuss and extract the knowledge of experts and the limited time that they are prepared to give to such a process. As with other components of scenario building (see the preceding case study), the input of experts is essential during this phase. To maximise the chance that the appropriate experts will attend, the workshops are limited to one day.

This limitation goes at the expense particularly of the probability-estimating session, which happens late in the day. Until then, the bulk of the available hours is spent agreeing on the current situation, considering the main variables likely to contribute to change in the humanitarian situation\(^{41}\), and on identifying the scenarios, their triggers\(^{42}\) and their humanitarian impact. Preparatory, bi-lateral meetings with some key experts can streamline the discussion, especially in enabling the facilitators to present an understanding of the current situation that is succinct and shared by participants and in identifying the main variables. However, even with extensive preparatory work (the facilitators have usually run through the whole exercise with at least one expert prior to the workshop), little time remains for the consideration of the relative probabilities of the various scenarios in one-day workshops.

For probabilities to be meaningful, the scenarios to which they refer should be so defined as to be exhaustive (any situation will be covered by one of them) and mutually exclusive (only one scenario is relevant in a specific situation). Only then will the scenario probabilities sum to one.

In workshops, those formal requirements are not always met. The consideration of the many variables of a crisis can lead to the development of scenarios that are not mutually exclusive. One main variable usually emerges as paramount (the conflict, the weather, etc.) and scenarios built around different assumptions as to the future behaviour of that variable (increased conflict, decreased conflict, continuing level of conflict, no conflict, etc.) will, by definition, be exhaustive as well as mutually exclusive. Workshop

\(^{41}\) Four categories of variables are considered: the current driver(s) of the crisis, possible future drivers, people’s resilience and vulnerability, and the in-country capacity to respond.

\(^{42}\) A trigger is an event that would cause, either alone or in conjunction with other triggers, the variable to change. E.g., peace talks fail; military forces are withdrawn from the northeast due to unrest elsewhere.
participants consider other variables (disease, natural disaster, etc.) significant. These are independent of the main variable, if not empirically, at least semantically. They are included in the scenario analysis as compounding factors. For planning or policy reasons, occasionally it is appropriate to develop a full scenario related to a variable other than the main one. In this case the scenarios are no longer mutually exclusive. The compounding-variables scenario then occurs in conjunction with one or more of the other scenarios. The sum of probabilities of a set of such scenarios may exceed one. For a suitable remedy, we return further below.

**Trigger probabilities**

Time constraints aside, workshop participants are challenged to define the probability of each scenario individually. Generally, this is done by involving experts who have good knowledge of the various triggers for a particular scenario. It is straightforward enough (if imprecise) for the relevant experts to attach a probability to each trigger that it will occur. The challenge lies in aggregating these individual trigger probabilities to an overall probability to realize that scenario. In fact, the challenge is double:

- The trigger probabilities are correlated. This means: Trigger B is more likely to materialize if Trigger A materializes than if A remains muted. The pattern of these correlations is imperfectly known, and under the time pressure of the workshop difficult to discern adequately.
- Not all triggers have the same capacity to bring about the scenario. For example, a scenario might have five triggers where the first is sufficient to trigger the scenario alone while at least three of the other four may have to occur to trigger the scenario in the absence of the first. Thus, there is a need to give each trigger a measure of importance, a weight that expresses its presumed effect.

As yet, ACAPS has not devised a robust way to aggregate these individual trigger probabilities and weights to produce an overall scenario probability. Even if an algorithm existed, it would not be credible in workshops typically filled with persons with few notions of probability. As illustrated below, the current method is to have experts work out the overall scenario probability intuitively, in the light of the various triggers. How individual experts process this information remains largely unknown.

**Probability ranges**

Ensuring the probabilities sum to one (or, more popularly, to 100 percent) is mitigated by using probability ranges, with a wide middle range and much narrower extreme ranges, as usually the probability of a large positive or negative change in the situation is small (less than 10%) whereas the probability of a moderate change is greater. In practice, the participants in most workshops choose two scenarios as being “likely” and a third as “possible”. A short explanation of “probability basics” is usually enough to convince participants to reclassify them as “possible” and “unlikely”. These, when we adopt the unequal probability ranges as in the illustration below, satisfies the need for the probabilities to sum to 100%. Ironically, few participants (and few facilitators) may realize that in information-theoretic terms the “possible” scenarios are uninformative (the probability of a possible scenario happening or not happening are the same). The
“unlikely” one is informative (the probability of its not happening is larger). But that matters less than the differences in the triggers and impacts of the scenarios. These are the stuff that informs policy.

[Sidebar:] Aggregating quantitative probability intervals

The probability that experts – in workshops or working alone – assign to scenarios needs to be clarified in one important aspect. The development of Area X or Vulnerable Group Y in the next Z months will happen only once, as a unique historic chain of events, demarcated by place, community and time. Therefore, we cannot think of probability as a likely frequency – the proportion of events in the total number of opportunities, such as the number of traffic accidents divided by the total number of road trips. Rather, the probability must be understood as our subjective degree of belief that the scenario will materialize within that defining frame. We are dealing with a population of one and sample of one – one crisis area and one upcoming period of time of interest that we are going to observe as a whole.

The degree of belief can be expressed as a number, such as from 0 to 100, with zero meaning “I am absolutely sure it will not happen”, 50 “completely unsure” and 100 “absolutely certain it will happen”, and any numbers in-between. With a quantitative degree of belief, one number – e.g., 90 for “I am almost positive” – makes sense, whereas intervals would not. However, most of us will be uncomfortable committing to one number and will prefer vague verbal expressions. Words like “possible” and “likely” can be understood to cover intervals of degrees of belief.

Occasionally, the experts assembled in a workshop have such accurate knowledge of the crisis that they can produce quantitative probabilities of a kind that will help in the scenario building. Suppose that it is known that in the half-year period just completed the violence claimed the lives of 1,000 civilians. The experts can be asked to submit estimates, individually and covertly, on a question like “Knowing the recent level of violence, what is your personal projection of civilian fatalities in the next six months? Write down the lowest number and highest number between which you are 80 percent sure the real number will fall”\(^{43}\).

Expert A submits “500 – 800” ; she is decidedly optimistic. B, counting on “500 – 2,000”, is undecided (her decrease and increase are the same, by a factor of 2 on both sides of 1,000). C, with “1,500 - 4,000”, believes in a significant deterioration. Etc.

How can such probabilistic intervals be aggregated? Park and Budescu (2015) discuss the research that has been conducted on this problem. They test various proposed methods with experimental datasets that hold both the experts’ estimates and the true values. They look for the best method of forming the aggregate interval – the interval should not be unnecessarily wide - correct in most circumstances, but also uninformative. Nor should it be too narrow - highly informative, but missing the true value most of the time.

They found the best method was the one that they named the Method of Quartiles. It has two instructions:

\(^{43}\) Soll and Klayman (2004) discuss other elicitation formats.
Humanitarian applications - Anticipatory analysis: Visualizing impact and probability

- Compute the aggregate lower bound as the 25\textsuperscript{th} percentile of the individual lower bounds.
- Compute the aggregate upper bound as the 75\textsuperscript{th} percentile of the individual upper bounds.

The more intuitive method of using medians produces intervals that the experiments have shown to be too narrow. Thus the more conservative quartiles.

We show the calculation in an Excel sheet with 10 experts with simulated interval estimates.

**Table 30: Aggregating experts' belief intervals - Simulated example**

<table>
<thead>
<tr>
<th>Expert No.</th>
<th>Expected fatalities next period</th>
<th>Relative change over previous-period fatalities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>1,000</td>
</tr>
<tr>
<td>2</td>
<td>800</td>
<td>1,000</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>1,000</td>
</tr>
<tr>
<td>4</td>
<td>800</td>
<td>1,500</td>
</tr>
<tr>
<td>5</td>
<td>1,000</td>
<td>1,200</td>
</tr>
<tr>
<td>6</td>
<td>1,000</td>
<td>1,200</td>
</tr>
<tr>
<td>7</td>
<td>1,000</td>
<td>1,500</td>
</tr>
<tr>
<td>8</td>
<td>1,000</td>
<td>2,000</td>
</tr>
<tr>
<td>9</td>
<td>1,250</td>
<td>1,500</td>
</tr>
<tr>
<td>10</td>
<td>1,500</td>
<td>2,000</td>
</tr>
</tbody>
</table>

25\textsuperscript{th} percentile of the lower bounds =PERCENTILE.EXC(R4C:R13C,0.25)

75\textsuperscript{th} percentile of the upper bounds =PERCENTILE.EXC(R4C:R13C,0.75)

Resulting aggregate interval 800 - 1,625 -20\% - + 63\%

The method is sufficiently simple for the facilitators to demonstrate the aggregation in the workshop plenary. The result is an interval that combines the beliefs of the experts. In this artificial example, the combined belief is that the relative change in violence will fall somewhere between a decrease by 20 percent and an increase of 63 percent.

The facilitators may spark a discussion around the result. The participants may agree that this is a “likely” range. However, in terms of humanitarian consequences, it is not a fruitful scenario. The range is too wide, comprising mild improvements and significant deterioration. The participants may agree to split the range into two scenarios. No. 1
would be a similar level of violence as before (say, a change from -20% to +20%), no.2 signifies a significant deterioration (+20% - +63%). Both scenarios may appear equally possible. In addition, two more scenarios can be defined. A significant improvement is understood as one with a decrease in fatalities of more than 20 percent. An extreme deterioration would bring an increase of more than 63 percent. The participants may consider both very unlikely.

The advantage of this method is not its precision, which is a pseudo-precision. Rather, it anchors the work of the participants in a common reference – the known and accepted fatality figure of the previous period. It allows them to express their personal beliefs in the evolution of the crisis in a format that captures both direction (increase vs. decrease) and certainty (width of the interval). There is a simple aggregation method that has some experimental grounding and can be demonstrated right away on the white board. It engages individual thinking among as many experts as there are in the room, as well as group debate.

To our knowledge, this method has not been tested in humanitarian information management workshops. It is noted here as a didactic opportunity.

An Illustration: Northeast Nigeria, October 2016 – June 2017

In September 2016, a set of scenarios considering possible developments of the conflict with Boko Haram in NE Nigeria were developed. For most humanitarian organisations, senior staff were located in the capital, Abuja, while those based in NE Nigeria were predominantly project staff. This presented a challenge in deciding the location of the workshop. It tends to be the more senior staff who are more interested in the process of scenario-building, as they are the ones involved in longer-range planning. Conversely, the project staff generally have a better understanding of the situation on the ground. This is especially true in Nigeria, where the turnover of international staff is high: few of those consulted for the scenario exercise had been in country for more than a year, and most between one and six months. Thus, two workshops were planned, the first to be held in Maiduguri (the main hub in NE Nigeria) on a Thursday and the second in Abuja the following Monday.

The scenario building followed the “chain of plausibility” approach (see footnote on page 150), in which the key variables that drive or contribute to change in the situation are identified and their relationships are mapped. The most interesting (or significant) variables are then selected, and various assumptions are made on how each chosen variable might change in the future. To make this process manageable, the “future” is usually limited to between six and twelve months. In this case, facilitators and Forum chair chose a time horizon of nine months, so as to include the whole of the dry season, during which fighting is more common. Three days spent in Abuja, and three in Maiduguri, prior to the workshops enabled the facilitators to meet key informants and gather specific background information to inform the scenarios, especially on the relevance and impact of the variables that drive change in the conflict.
Fourteen participants attended the first workshop in Maiduguri: two from the government, five from UN agencies, six from NOGs and one from the ICRC. Although most were Nigerian and had a good understanding of the Boko Haram conflict, few had sufficient confidence or willingness to engage in the discussions. Experience shows that the most productive workshops are those attended by participants who regularly engage in analysis and strategic thinking and have a good knowledge of the complexities of the crisis and context. While participants in Maiduguri had much relevant knowledge they were not used to analytical or strategic thinking. This workshop resulted in the development of six mini-scenarios. The ACAPS facilitators then re-worked these mini-scenarios, removing two and slightly re-wording the others.

The four scenarios retained were presented and further developed at the workshop in Abuja. Their defining features were:

1. Low-level conflict continues while increasing numbers of people return.
2. Security improves significantly (although high profile terror attacks also occur) resulting in multiple displacement flows.
3. Security deteriorates significantly across the region; famine spreads.
4. A peace settlement is negotiated, resulting in large-scale returns.

Most of the time was devoted to discussing triggers and impacts. Towards the end, there was a quick 45-minute discussion of the probability of each scenario. Various participants suggested (verbally defined) probability ranges for each scenario and were then challenged by the facilitator (and other participants) to justify their suggestions. This process, whilst largely unstructured, enabled the combined knowledge and experience in the room to gauge probability and impact. It let the facilitators understand the various arguments for and against. It became apparent that equal numbers of participants considered scenario 1 to be as ‘likely’ as scenario 2 while there was general agreement that scenario 4 was ‘highly unlikely’ and 3 ‘unlikely’ but not as unlikely as 4. This was understandable as no. 1 was essentially a maintenance of the status quo while no. 2 was a continuation of the current trend of the conflict.

This failure to reach complete agreement is common and, as is ACAPS practice, the facilitators subsequently determine the final probability rankings based on the workshop discussions, other bilateral meetings and the need to avoid classifying more than one of a number of mutually exclusive scenario as ‘likely’ or ‘highly likely’. In this case, the ACAPS re-classified both scenario 1 & 2 as ‘possible’ as it is illogical to have more than one ‘likely’ scenario in a setting with solely mutually exclusive scenarios.

Scenario 4, while universally considered highly improbable, was included as it was thought that such a scenario would result in huge humanitarian needs in areas of return while the international humanitarian community would be cautious in scaling up to meet the need.

The fifteen participants represented nine organisations (NGOs, OCHA and the government). Most of them were junior staff, and none of the Maiduguri participants joined. The workshop suffered also from a lack of specialist knowledge in many areas. As a result, in the days after the workshop the facilitators approached other experts to
Humanitarian applications - Anticipatory analysis: Visualizing impact and probability

discuss a number of specific triggers and impacts. By synthesizing workshop findings and complementary expert opinions, the facilitators were able to posit for each scenario a probability range, a likely direction of the impact and the expected duration of the crisis. The figure exemplifies the visual language expressing probability and impact.

**Figure 33: Visual scales of probability and impact**

![Scenario 3: Widespread insecurity, famine](image)

Duration of crisis: More than 2 years

Overview: At least one of the insurgent factions changes tactics and regains popular support in some rural areas, increasing in strength. The conflict escalates, and spreads again into other states. Displacement increases significantly. Across the northeast, humanitarian access reduces as does the state’s, already limited, capacity to provide services.

**Source:** Nigeria INGO Forum and Acaps (2016:2).

**Precision vs. plausibility**

Assigning probabilities to scenarios is not an exact science. By using a verbal scale (‘Highly likely’, ‘Likely’, etc.) with each level signifying a quantitative range and avoiding the use of percentages, ACAPS aims to give some sense of relative probability of each scenario materializing while avoiding the appearance of precision. It is, after all, important for humanitarians to consider all scenarios during strategy and program development and not just those most probable. The colored probability bars, immediately above the impact rating bars, are a visually powerful means to summarize these two key attributes for the users’ easy perception.

The use of a probability scale of equal ranges has proven difficult. Theory says that the sum of the probabilities of these mutually exclusive scenarios cannot exceed 100 percent. If we had to quantify the probability of each of the NE Nigeria scenarios with point estimates rather than ranges, it would be roughly 1: 40%, 2: 40%, 3: 15% and 4: 5%. Using the equal range scale below, as was ACAPS practice, resulted in scenarios 1 and 2 being ‘unlikely’, and 3 and 4 ‘highly unlikely’. This did not reflect the views of the workshop participants and is not particularly helpful to those reading the report – as everyone agreed that the most likely future would see either a continuation of the status quo (scenario 1) or a slight improvement in security (scenario 2). Having them both categorized as ‘unlikely’ is therefore misleading; so the decision was taken by ACAPS to show the relative probabilities by having both 1 and 2 as possible, 3 as unlikely, and 4 as highly unlikely, despite their not summing to 100%.
Medow and Lucey (2011) proposed a scale based on unequal probability ranges (see pages 99 sqq. for an extensive discussion). It offers more flexibility in view of commonly felt large uncertainty, and therefore better consistency of the overall probability distribution of all the scenarios. Using these unequal ranges enables the chosen probabilities of the NE Nigeria scenarios to satisfy the mathematical requirement of summing to no more than 100%.

Table 31: Equal vs. unequal probability ranges

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Equal ranges</th>
<th>Unequal ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly likely</td>
<td>81 – 100%</td>
<td>90 – 100%</td>
</tr>
<tr>
<td>Likely</td>
<td>61 – 80%</td>
<td>67 – 90%</td>
</tr>
<tr>
<td>Possible</td>
<td>41 – 60%</td>
<td>34 – 66%</td>
</tr>
<tr>
<td>Unlikely</td>
<td>21 – 40%</td>
<td>11 – 33%</td>
</tr>
<tr>
<td>Very unlikely</td>
<td>0 – 20%</td>
<td>0 – 10%</td>
</tr>
</tbody>
</table>

In a set-up with four scenarios, two rated “possible” and one each “unlikely” and “very unlikely” form a feasible combination. Three “possible” and one “very unlikely” scenarios would not be a good combination, probability-wise, because $3 \times 34\% + 0\%$ is already $> 100\%$. The Medow-Lucey scale encourages both flexibility and discipline. Thus, while it may happen that the workshop participants consider all four scenarios “equally possible”, that would be a purely conversational statement as opposed to one based on a quantitative interpretation. The sum of probabilities, in percentages, would then be at least $4 \times 34\% = 132\%$. What the participants really mean by “all four are equally possible” is: “all four are equally unlikely”, as in $4 \times 25\% = 100\%$. The scale disciplines talk about probability.

Finally, we need to briefly reconsider the extent to which scenarios are mutually exclusive. Two scenarios are mutually exclusive if none of the combinations of the attributes that define Scenario A overlaps with the attribute space defining Scenario B. In the Nigeria study, the level of security/insecurity differentiated the four scenarios (negotiated settlement / increasing security without a settlement / continued low-level conflict / widespread insecurity). Other attributes were accidental (returns / displacements; famine), pointing to major consequences of the particular security levels.

Since the descriptors of the defining attributes are vague, mutual exclusivity is a scientific ideal, but in practice is a matter of degree. Obviously, there are transition states, such as between continued low-level conflict and widespread insecurity. In this case, a range of insecurity states may qualify as equally credible realizations of either scenario. A pessimistic observer might therefore consider the probability that NE Nigeria experiences either continued low-level conflict (Scenario 1) or widespread insecurity (Scenario 3). These scenarios were considered “possible” (34 – 66%), respectively “unlikely” (11 – 33%). The probability that either of them materializes is $= 1 - (1 - \text{prob1}) \times (1 - \text{prob3})$. Using the range bounds, we obtain the interval $[1 - (1 - 0.34)*(1 - 0.11), 1 - (1 - 0.66)*(1 - 0.33)] = [0.41, 0.77]$. 
The probability that either of the two scenarios materializes is 41 – 77 percent. This result, retranslated to verbal expressions, is in the grey zone between “possible” and “likely”. This seems plausible. Note that this range is wider (36 percent) than either of the individual probability ranges, which is what we intuitively expect. Conversely, an optimistic observer would want to know the probability that either increasing security (Scenario 2, possible) or a negotiated settlement (Scenario 4, highly unlikely) happen. The probability range for that, calculated by the same method, is 34 – 69 percent, which is almost entirely within the quantitative definition of “possible”. Both results suggest that the method of probability ranges offers both the consistency and flexibility that a rational discussion of scenarios requires.

**Lessons learned**

Assigning probabilities to scenarios is beset with challenges, for some of which good solutions exists whereas others are resolved in ad-hoc and arbitrary ways. In particular, there is no method for aggregating trigger probabilities other than the personal intuition of the experts.

At the scenario level (as opposed to triggers), probabilities scales based on unequal ranges – preferably that proposed by Medow and Lucey (see details above) – produce probabilities that sum to one if the scenarios are mutually exclusive. If the experts are comfortable expressing probable outcomes (e.g., fatalities in the next six months) by a numeric range, the method of quartiles, proposed by Park and Budescu (see sidebar within this case study), comfortably produces an aggregate range, easy to demonstrate in a workshop.

Workshop dynamics are important throughout. Frequently, not enough time is left to discuss probabilities, and the understanding of probability among participants may be so limited as to frustrate discussion. Strong guidance in the common workshop, and separate interviews with individual experts are recommended.

**5.11. Prescriptive analysis: Experts vs. decision-makers**

Prescriptive analysis talks about what could and should be done, in view of the preceding analysis, and considering the options that are appropriate, high priority, and feasible. ACAPS does not normally engage in prescriptive advice. Yet, just as medical doctors diagnose and prescribe, a study on expert judgment would be incomplete without a look at the function of experts in prescriptive analysis.

**An illustration – Strategy change in Somalia, 2010-11**

For illustration, we turn to the famine that ravaged Somalia in 2011, one of a series of disasters within a persistent complex emergency. The famine, which claimed more than 200,000 lives, provoked intense self-reflection in the humanitarian community. The use and quality of expert judgment in assessment and response were not an explicit rubric in any of the retrospective analyses that we read. In this long-standing humanitarian theater, organizations and individuals with profound area and technical expertise were numerous; how decision-making in this crisis was infused with expert judgment transpires only implicitly. De facto, both major providers, the United States Agency for International Development (USAID)-funded Famine Early Warning Systems Network
Other themes dominated the retrospective. The inadequacy of response analysis, particularly the paucity of contingency plans, was a major one. Ironically, challenges of response analysis and planning had been extensively debated among agencies engaged in Somalia, particularly between the Food Security and Nutrition Clusters, before the disaster (FAO and WFP 2011a, FAO and WFP 2011b). Although chronic insecurity forced much of the humanitarian action to be managed remotely, Somalia by that time was an information-rich environment. The early-warning system functioned; the risk of impending famine was correctly assessed and repeatedly communicated (Haan, Devereux et al. 2012).

However, the international community did not significantly scale up relief until the United Nations declared a famine, in July 2011, when it had fully developed. Uncertainty compounded by lack of access, under-subscribed funding requests, varying beliefs in the feasibility of strategies, and coordination burdens stood in the way of a timely response. Programs in South Central Somalia were “adjusted” in minor degrees in late 2010 and were “adapted” to life-saving objectives, although without increased operational capacity, in the first half of 2011. Only the declaration of famine in July caused donor, international agencies and implementers to re-align their strategies for a massive “transformative” response (Hobbs, Gordon et al. 2012).

The western public was presented with a famine as the result of a severe drought. In reality, a series of factors cooperated in curtailing access to food and in withholding timely relief. Hostility to aid agencies by the Al-Shabaab armed movement as well as elements of the Transitional Government and the ensuing withdrawal of the World Food Program and of the ICRC removed capacity to scale up relief. US anti-terrorism laws interfered with relief options and with the flow of remittances in the contracting economy (Maxwell and Fitzpatrick 2012). By the time the formal declaration of a famine rapidly mobilized resources, WFP’s capacity to implement had atrophied. With massive food aid reduced as the normal option, non-traditional responses were scaled up, in the form of cash assistance, cash-for-work in agricultural rehabilitation, and nutrition programs. For the interested reader, Maxwell and Fitzpatrick, op.cit., and the series of studies in the same issue of Global Food Security provide detailed chronology and analysis.

Information-rich, interpretation-challenged

Somalia may have been a data-rich environment, but the interpretation of so much data percolating through many organizations was problematic, and working out consensus interpretations remained challenging. Not information scarcity, but sense-making was particularly limiting. In the words of Haan et.al. (op.cit., 74), “there has rarely – if ever – been a crisis and a response that was so apparently rich in data, while so lacking in...
any human sense of what was happening on the ground among the affected population groups due to lack of humanitarian access”.

For us the question of interest is the role of expert judgment in the situation and response analyses that guided the famine relief in 2011. There are no indications that expert judgment in the sense of substituting estimates for observational data was problematic or conflictual. However, the causal models that ultimately informed beliefs of effective interventions were not well integrated across agencies and disciplines. In workshop reports and discussion papers, complaints abounded that frameworks were disconnected, particularly between food security and nutrition (FAO and WFP 2011a:35). At the institutional level, the WFP’s own assessment was that response analysis was still weak, and guidance was still in the process of being developed (FAO and WFP 2011b:7). The FAO Response Analysis Framework (RAF) had been introduced to Somalia, with “some common problem analysis across the nutrition, food assistance and agriculture and livelihoods clusters, but it was only the latter cluster that really internalised and used the RAF in the development of its 2011 CAP response plan” (ibd., p.9). Prioritization was insufficient, notably because coordination and consensus were difficult to achieve. To illustrate: the possible switch from food aid to cash assistance was debated in “some 15 different coordination forums” (Haan et.al., op.cit.:76).

One comes away with the impression that technical expertise was not in short supply, but that technical and policy experts missed a bridge between them. As the reviewers of the early warning process put it:

“In spite of recent success, early warning systems must also continue to improve. Efforts to better integrate market data, political/conflict analysis, livelihoods information, and nutrition analysis should continue, as should the further improvement of scenario development techniques. Most importantly, these systems must further refine decision support tools (e.g., written reporting, briefings, and web content) to better communicate complex technical analysis and address decision-maker concerns with uncertainty. This should include greater sensitization of decision-makers to the newest version of the IPC [Integrated Phase Classification] and to scenario development” (Hillbruner and Moloney 2012:27).

It is hard to see, however, how more or better policy experts thrown at the response planning could have released the constraints that the belligerents’ behavior and the agencies’ reluctance in the face of Al-Shabaab’s designation as a terrorist organization were imposing. Hobbs et.al. (op.cit., 55) correctly observed that “in dynamics of enduring crisis, where normal conditions are characterized by significant levels of humanitarian need, it is difficult for humanitarian actors and donors to determine when a radically altered response should be initiated”. Their conclusion seems self-evident: “This underlines the importance of supporting on-going consensus building and response planning processes in situations of protracted crisis” (ibd). But it is doubtful that an organizational field as complex as the one in and about Somalia is capable of sustaining processes that deal not only with the normal business, but permanently also
with a range of extreme contingencies that might require “transformative action”. As long as these seem distant risks, capacities – including expert judgment and other ingredients of response planning – are likely to be kept at a base level.

Response analysis beyond Somalia

Less than two years later, Humanitarian Practice Network (HPN) Network Paper No. 73, “Response analysis and response choice in food security crises: a roadmap” (Maxwell, Stobaugh et al. 2013), reviewed the general case for response analysis. Response analysis is defined as “the analytical process by which the objectives and modality of programme response options in an emergency are determined, and potentially harmful impacts are minimized” (p.3). Essentially, this is a work in terms of policy expertise – a panorama of methodologies of generating and selecting options for such response. A good part of the policy expertise consists of knowing where to find the tools of technical expertise. The annex details a large inventory of analysis tools and decision trees. Inspired by the Somalia experience, the authors emphasize that “response analysis is an ongoing process, not a single step” (p.25). The tone is optimistic because “donor resources are more flexible now and agency capacity is rapidly evolving” (p.3).

The key insight springs from an ordering of options that the analysis has collected. The authors distinguish them as first-, second-, and third-order options. First-order options are strategic; “this first order of decision-making tends to set the boundaries of the response, but it does little to fill in the details”. Second-order options define “the modality in which to achieve the general objective set by the first-order decision” (p.4). Third-order choices expand greater detail. In food security, for example, choosing between conditional or unconditional aid is an important third-order decision. Contrary to what the numbering may suggest, the ordered options are not strictly hierarchical. “Many of these details weigh heavily in the response decision-making process, even to the point of solely determining the overall objective or modality of a programme. Thus, these choices are still considered a part of response analysis, rather than programme design (albeit the line is fuzzy between the two)” (p.5). Although the referenced decision trees are strictly hierarchical, by and large response analysis proceeds through iterative processes in which ends and means enable and constrain each other mutually.

The interacting orders of options and choices are further opened by cross-cutting considerations that must be given due attention in response analysis and planning. They involve “various forms of risk assessment” (p.5), from the distortion of markets, to environmental impacts and security risks, and more. In the expert judgment perspective, it is plausibly the realm of cross-cutting considerations that creates the greatest demand for technical expertise. It may also be responsible for unresolved and unresolvable issues, which, in turn, further stimulate expert work, both of the policy and technical kind. For example, “while much emphasis has gone into improving targeting in recent years, considerable controversy still exists about the most appropriate form of food assistance targeting (administrative, community-based or self-targeting)” (p.5). The

---

demand for expertise is counteracted by the perceived excessive complexity of some of the tools that experts have created, the effective application of which would necessitate even more support by experts:

“While programme teams are often aware of at least some tools, these tools are not frequently used. Several reasons were reported to explain why this is the case. Firstly, there are too many to choose from, and it is not always clear what tool is used for what task or decision. Secondly, and probably most importantly, existing tools are seen by many practitioners as being too complex, too time-consuming and requiring too much technical expertise. Thirdly, none of the tools really maps out the overall process of response analysis; rather, they tend to be specific to one particular kind of decision” (p.19).

The demand for expertise, it seems, is both self-propelling and self-limiting.

**Experts cannot ensure coordinated decision-making**

The Somalia famine in 2011 was a dramatic and internationally prominent event. Whatever history has to say about the adequacy of the response, Somalia was not the stage for humanitarian expert judgment to be particularly triumphant or particularly failing. The early warning systems worked; this indicates that good judgment was available and offered, if not widely heard. The crisis perhaps was driven more by what others have called the “breakdown of coordinated decision making in distributed systems” (Bearman, Paletz et al. 2010); if so, this is beyond the scope of our study. Of interest here is the fact that the crisis experience was analyzed for a wider audience, that the analyses were performed by persons who themselves are experts, and that their reports touch upon some of the potential and limitations of expert judgment. There seems to be an inbuilt mechanism to make expert judgment self-reflective. This is likely true of the prescriptive as well as of the other kinds.

**Lessons learned**

The crisis demonstrated the limited impact of expert judgment in the face of decision making borne by power/violence, resources and international law. Yet, it also triggered waves of collective reflection that, among other effects, have stimulated expert judgment methods for better response planning.
6. Conclusions and outlook

**History and present**

“Expert opinion”, in the words of a classic author in this field, “is cheap, plentiful, and virtually inexhaustible” (Cooke 1991:op.cit., 3). This sweeping claim is exaggerated. Surely, experts can be better, faster or cheaper than other sources of knowledge and information. If better, they may still not be good enough; if faster, still slower than would be useful; if cheaper, still expensive. It all depends. And good judgment may be plentiful in well-resourced organizations, but not inexhaustible.

Cooke’s optimism stands in a line of expert judgment that came to maturity in the US nuclear power industry. Ours is a study of expert judgment in the humanitarian domain. Towards the end it is befitting to briefly return to historical origins and to recall differences in the ways experts function in those two vastly different environments. This historic aside is important in order to appreciate the potential and limits of expert judgment in the humanitarian world.

In nuclear power, most of the experts were engineers; they would address questions that were posed to groups of them simultaneously, calling for answers that were largely quantifiable. The risks of nuclear accidents was clearly present, and the business of safety experts was dead-serious. Yet the organizational environment of their work was placid and stable, the result of dense regulation and public debate. The most serious accident in the USA – the partial reactor meltdown at the Three Mile Island plant in 1979 – was managed competently and calmly. The response vindicated both expert quality and work environment, in contrast to the political and corporate environments that enabled Chernobyl and Fukushima.

There are aspects of the classic expert judgment methodology that transfer to the humanitarian domain almost in their entirety. The imperative to clearly define expectations at three levels – overall goal, broad question areas, specific questions – is a good example. Other aspects are significantly modified as we move from the historic context to the humanitarian. Thus, as some of our case studies show, humanitarian experts turn over faster, and analysts are thus limited to elicitation formats that are robust to changes of place and personnel, or settle for lower quality.

Regardless, experts fill data gaps when judgment replaces direct observation and measurement (because these are unavailable or of insufficient quality). Moreover, experts have superior skills to process data that arrive from observations and measurements, made by them or by others. They have methods to deal with the uncertainty in judgments as well as in observational data. There is an established science of expert judgment; the methods of formulating questions, getting answers and combining them are tested.

At this point the commonalities end. The humanitarian environment is turbulent. The turbulence has consequences for the functioning of expert judgment. The roles of decision-maker, analyst and expert are less differentiated. The individuals that fill them
move frequently; personnel turnover is high. Technical and policy expert skills are required in stronger combination, as inter-sectoral competence or at least empathy. Qualitative and quantitative information are harder to separate; yet mixed methods have not widely penetrated humanitarian analysis. The bases of decisions grow out of sequential contributions, rather than from the consensus of experts convened simultaneously. The collection and aggregation of judgments have to overcome language, access, conceptual and skill barriers. Local knowledge is indispensable, making the key informant as important as the technician. The actual use of expert judgments in decisions is rarely documented. Limited use of formal methods, time pressure and absence of legal incentives favor pragmatic interaction styles.

Of course, those are tendencies, not absolutes. The habit of updating beliefs in the light of new evidence is natural in the humanitarian world. Humanitarian researchers, information managers and other experts should therefore be highly receptive to the Bayesian revolution and to methods and technologies that favor regular, even near-real-time updating. At present, there are few, if any, explicit Bayesians among them. The culture of the situation report does not always clearly mark the uniquely new, the old and the redundant, all of which have their positive functions.

**ACAPS’ experience**

For those many reasons, the fit between formalized methodologies of expert judgment and humanitarian analysis is partial. But it is evolving, as ACAPS’ work demonstrates. There has been cumulative methodological progress, informed, sustained and replicated by many partners. There are more practices known to work, or known to work poorly or not at all. At the same time, as the case studies demonstrate, all field situations demand improvisation – adaptations that were not anticipated in earlier missions or in lessons-learned exercises. Improvisation guided by persons with a firm grip on the toolbox – in other words: experts – is superior to mere tinkering, but succeeds only in a combination of openness for novelty and disciplined coordination.

There remain obstacles, some of which are difficult to step over. ACAPS is at a low point on the learning curve with probabilistic models. As noted, assigning probabilities to trigger events and then aggregating them for the scenarios is an unresolved challenge, in terms of algorithms as well as of workshop didactics. Expertise in mixed qualitative-quantitative information analysis is another area that awaits fresh ideas, without losing our way in the sprawling thicket of the methodological literature.

**What comes next?**

These observations take us into some futuristic speculation. It is a safe bet that the development of Web-based applications and the Big Data revolution will unsettle the ways experts are recruited and work in the humanitarian world. Pressures for “evidence-based policies” and “value for money” work in the same direction. It is harder to predict how specifically this will happen.

Plausibly, innovations in expert judgment will advance at an uneven pace. Technologies that generate and process massive data of quantitative and categorical nature may be adopted with relative ease. There will be selection issues – segments of
populations in need that are not connected to the data-generating process or are misrepresented by it. And there will be experts with specialized methods to mitigate such risks, alongside others who deal with spam and intrusion.

We may also see innovations in dealing with qualitative expert judgment, such as in Web- or Intranet-based “argumentative Delphi”, a technique that generates, evaluates and combines arguments and counter-arguments from a potentially large group of participants (Seker 2015). Technique and infrastructure will be able to connect novel coalitions of experts, key informants and even persons directly at risk in the evaluation of needs, environments and response options. There may be a continuum from formats still controlled by experts to more self-organized processes, an inkling of which we saw among refugees counseling each other over mobile phones as they adjusted itineraries on the Balkan route in 2015.

There will also be counter-movements, calls for experts to guide slower deliberations, using smaller datasets or none, ethnographic methods, multi-round revisions of accustomed conceptual distinctions, thinking about the bigger questions of humanitarian action. Even a “Slow Delphi” has been proposed (Poirier and Robinson 2014), an offshoot from the cultural resistance known as the “Slow Movement” (Pels 2003, Honoré 2009). An early manifestation was the “Do no harm” movement started by Mary Anderson (1999).

**Expert judgment as sense-making**

Going slow is not an option under life-saving imperatives. Still the disconnect between experts working to accelerate humanitarian intelligence and those whom agencies seek out for their detached and reflective views can be reconciled. The concept of “organizational sense-making” straddles both cultures. This is not the place to retrace its long tradition, much of it owed to the work of Karl Weick (1988, 2012). In brief, “sensemaking involves turning circumstances into a situation that is comprehended explicitly in words and that serves as a springboard into action” (Weick, Sutcliffe et al. 2005). Understanding, talking and doing come together.

How do experts help humanitarian organizations make sense of what’s going on, and ideally also of what the organizations can do and should do?

We believe experts are effective when they work with analysts and decision-makers to balance accuracy and plausibility. Expert judgments narrow gaps in data and result in superior analysis. This leads to more accurate estimates, which, in theory, enhances decision quality. By contrast, decision-makers confront the “What’s the story here?” and “Now what shall we do?” questions. These beg for, not accurate, but plausible, answers. While the expert accurately assesses scarce data (and generates some more), the decision-maker interprets an abundance of information into actionable knowledge (see Weick et al., op. cit., 415).

What about the analyst in this scheme? It is tempting to assume that his role is essentially one of shuttle diplomacy between accuracy and plausibility concerns. This is an important function, but not exclusively his. When experts become involved over
longer periods and with deeper insight into the workings of the organization, they themselves reduce the distance between accuracy and plausibility. As they move from description and explanation to anticipation and prescription, individual analyses increasingly become shared analyses. Tan and Chan (2016) show how the character of expert work changes in the transition. The abstraction level and the scope of the analyses increase; the granularity of the data decreases in favor of summaries and indices; techniques that we would rate as objective algorithms give way to more immediate human reasoning. The experts communicate better in the concepts of the organization’s own domain knowledge. Decision-makers, analysts and experts find more common ground.

The reader may find this conclusion persuasive - or not. He/she may feel that this study has been long on accuracy concerns, and rather short on plausibility, paying too much attention to experts and not enough to decision-makers. This may be so. We have undertaken the study in view of the scarcity, in fact almost total absence, of expert judgment literature situated in the humanitarian field. We believe that despite its unbalanced eclecticism it can contribute to effective cooperation among decision-makers, analysts and experts.
Appendices

ANNEX A

Excel VBA code for the triangular distribution

Purpose

The triangular distribution is an important probability distribution in simulation studies and in expert judgment. It is often used when little detail is known about the distribution, but plausible assumptions or (fairly) reliable observations can be made about minimum, mode and maximum. In this study we refer to its use in the estimation of the populations in the 270 sub-districts in Syria in 2015. As noted,

The distribution is popular particularly among risk management applications. It has three parameters that the expert (or key informant) supplies in response to:

- “What is your best estimate?”
- “What is the minimum below which you will not go?”
- “What is the maximum above which you will not go?”

The Wikipedia article gives a good overview about the statistical properties of the distribution (Wikipedia 2015). On pages 68 - 74 of this study, we discuss technicalities of drawing random variates from the triangles specified by multiple key informants. This is done for the purpose of creating a combined distribution and then, if called for, summing sub-district estimates to higher-level units, with point estimates and confidence intervals.

Ideally, an ambitious analyst should be able to perform the necessary operations all by herself, using a spreadsheet application. However, Excel does not offer the triangular distribution in its standard statistical formula portfolio. The gap is easy to close with a few snippets of Visual Basic for Applications (VBA) code stored in a module in the background of the Excel workbook. The code enables the user to call three triangular distribution-related functions.

The purpose of this appendix is to explain the three functions, demonstrate the placement of the code in a module, and calculate an example of each of them.
Appendices - Excel VBA code for the triangular distribution

Functions

**Probability density function (PDF)**

The probability density function returns, for a given point $x_0$ of the base $[a, b]$ of the triangle, the height at that point. In other terminology, it is the marginal increase of the probability at this point, which is the probability of $x$ falling within the infinitesimal interval $[x_0, x_0 + dx]$.

![Diagram of triangular distribution](image)

The area of the triangle is, by definition, $= 1$ (the probability that a randomly drawn value falls on the interval is 100 percent); thus the triangle at its highest point has a density $PD = 2 / (b - a)$. This is so because, by elementary geometry, the triangle area works out as $(1/2) \times (b - a) \times \left(\frac{2}{b - a}\right) = 1$.

The probability density of the triangular function is trivially computed on notions of similar triangles. The formula (see further below) must also reflect that outside the triangle base the density is zero. In practice, it may be of use (rarely) in comparing point densities in two or more overlapping triangles and (more usefully) for graphing multiple overlaid triangles as Excel connected-point graphs.

**Cumulative distribution function (CDF)**

The cumulative distribution returns, for any given point $x_0$, the probability that a randomly drawn value $x$ is $\leq x_0$. In geometric expression, it is equal to the area of the blue triangle over the base $[a, x_0]$. In this diagram, $x_0$ is smaller than $c$. If it were larger than $c$, we would color the entire area under the triangle from $a$ to $x_0$, i.e. the triangle over $[a, c]$, plus the trapezoid over $[c, x_0]$. The calculation again follows from elementary geometry. The cumulative probability increases non-linearly, i.e. proportionate to the square of $(x - a)$ for $x < c$, and thence onward to the square root of $(b - x)$.
Therefore, it is helpful to represent the CDF such that the cumulative distribution corresponds to a point rather than an area. This form maps the cumulative probability onto the interval [0, 1], in other words onto the full range of probabilities from zero 100 percent. This is ideal for the understanding of the third function, the inverse cumulative.

**Inverse cumulative distribution function (ICDF)**

This function responds to the question: “If I start by setting the probability \( p \) that \( x \leq x_0 \), where on \([a, b]\) is \( x_0 \)?” It is the “opposite”, in a manner of speaking, or the inverse of the CDF. For example, if our key informant for the sub-district population gives 10,000 as her best estimate, 5,000 as the minimum, and 20,000 as the maximum, she defines a triangular distribution. What is the median \( x_0 = x_{\text{median}} \), written as \( p(x \leq x_0) = 0.5 \) ?

The inverse is the most important of the three. It is the workhorse for the simulations that aggregate the triangular functions by several experts. The simulations essentially rely on large numbers of (ideally 1,000) draws of \( p \) from the uniform distribution on \([0, 1]\).

In other words, when, for example, 1,000 uniformly distributed random numbers are generated using the Excel function `RAND()`, the user-defined function `InverseTriangular()` will transform them into \( x \)'s such that “very few” of these are close to the corners \( a \) and \( b \), and increasingly more as we move to the mode \( c \) – just as the intuition of “minimum, best estimate, maximum” suggests.

**Code and placement in a VBA module**

In order to access user-defined functions, we need an Excel workbook saved in the macro-enabling extension `.xlsm`, such as, for demonstration purposes, `TriangularDistributionFunctions.xlsm`.

Pressing the keys Alt + F11 gives access to the “Visual Basic for Applications” (VBA) window. In Insert menu, we choose “Module” and optionally rename, in the Properties Box, Module1 as “Triangular” (to exhibit its purpose of hosting the code for these functions; later the reader may add other modules for other purposes). We copy-paste the following code into the lower right pane, then save:
Option Explicit

Function InverseTriangular(CumulProb As Double, Minimum As Double, Mode As Double, Maximum As Double) As Double

Dim LowerRange As Double, HigherRange As Double, TotalRange As Double
Application.Volatile

LowerRange = Mode - Minimum
HigherRange = Maximum - Mode
TotalRange = Maximum - Minimum

If CumulProb < 0 Or CumulProb > 1 Then
    InverseTriangular = ""  
ElseIf CumulProb < (LowerRange / TotalRange) Then
    InverseTriangular = Minimum + Sqr(CumulProb * LowerRange * TotalRange)
ElseIf CumulProb >= (LowerRange / TotalRange) Then
    InverseTriangular = Maximum - Sqr((1 - CumulProb) * HigherRange * TotalRange)
Else
    End If
End Function

Function TriangularPDF(InputVar As Double, Minimum As Double, Mode As Double, Maximum As Double) As Double

Dim LowerRange As Double, HigherRange As Double, TotalRange As Double
Application.Volatile

LowerRange = Mode - Minimum
HigherRange = Maximum - Mode
TotalRange = Maximum - Minimum

If InputVar < Minimum Or InputVar > Maximum Then
    TriangularPDF = 0
ElseIf InputVar >= Minimum And InputVar <= Mode Then
    TriangularPDF = 2 * (InputVar - Minimum) / (TotalRange * LowerRange)
ElseIf InputVar >= Mode And InputVar <= Maximum Then
    TriangularPDF = 2 * (Maximum - InputVar) / (TotalRange * HigherRange)
Else
    End If
End Function

Function TriangularCDF(InputVar As Double, Minimum As Double, Mode As Double, Maximum As Double) As Double

Dim LowerRange As Double, HigherRange As Double, TotalRange As Double
Application.Volatile

LowerRange = Mode - Minimum
HigherRange = Maximum - Mode
TotalRange = Maximum - Minimum

If InputVar < Minimum Or InputVar > Maximum Then
    TriangularCDF = 0
ElseIf InputVar >= Minimum And InputVar <= Mode Then
    TriangularCDF = 2 * (InputVar - Minimum) / (TotalRange * LowerRange)
ElseIf InputVar >= Mode And InputVar <= Maximum Then
    TriangularCDF = 2 * (Maximum - InputVar) / (TotalRange * HigherRange)
Else
    End If
End Function
Appendices - Excel VBA code for the triangular distribution

    Total Range  As  Double
    Application. Volatile

    Lower Range  = Mode - Minimum
    Higher Range = Maximum - Mode
    Total Range  = Maximum - Minimum

    If InputVar < Minimum Then
        TriangularCDF = 0
    ElseIf InputVar >= Minimum And InputVar <= Mode Then
        TriangularCDF = (InputVar - Minimum) ^ 2 / (TotalRange * LowerRange)
    ElseIf InputVar > Mode And InputVar <= Maximum Then
        TriangularCDF = 1 - (Maximum - InputVar) ^ 2 / (TotalRange * HigherRange)
    Else
        TriangularCDF = 1
    End If
End Function

Note the "_" (underscore) character ending some of the code lines. The underscore is the line continuation character in VBA. It makes sure that when the code is copied from a .pdf-formatted document (like the one you are looking at) Excel will not interpret the line break sign as the premature end of a command line and abort the execution.

This is a screenshot of a segment of the VBA window:
Appendices - Excel VBA code for the triangular distribution

Figure 35: Screenshot of an MS Excel VBA module

Formulas and examples
We calculate an example of each of the three functions. The input values are entered in column 2 in the blue area.
Appendices - Excel VBA code for the triangular distribution

Table 32: Examples of calculated functions of the triangular distribution

<table>
<thead>
<tr>
<th>Parameters and values:</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>EstMinimum</td>
<td>1000</td>
<td>estimated by the expert.</td>
<td></td>
</tr>
<tr>
<td>EstBest</td>
<td>1300</td>
<td>estimated by the expert.</td>
<td></td>
</tr>
<tr>
<td>EstMaximum</td>
<td>2000</td>
<td>estimated by the expert.</td>
<td></td>
</tr>
<tr>
<td>x0</td>
<td>1500</td>
<td>chosen by the analyst (example).</td>
<td></td>
</tr>
<tr>
<td>prob</td>
<td>0.5</td>
<td>chosen by the analyst (example), or a random draw using =RANDBETWEEN(0,1)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calculated functions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 Probability density</td>
</tr>
<tr>
<td>13 Cumulative probability</td>
</tr>
<tr>
<td>14 Inverse cumulative</td>
</tr>
</tbody>
</table>

with their formulas:

| 16 Probability density | =TriangularPDF(x0,EstMinimum,EstBest,EstMaximum) |
| 17 Cumulative probability | =TriangularCDF(x0,EstMinimum,EstBest,EstMaximum) |
| 18 Inverse cumulative | =InverseTriangular(prob,EstMinimum,EstBest,EstMaximum) |

To facilitate a generic notation – in other words, to avoid writing numbers into the formulas that may change as the reader tries out different examples –, we name the cells with values in R2C2:R8C2 by their corresponding terms in column 1.

The calculated values of the functions appear in R12C2:R14C2. Because in the example prob = 0.5, 1,408.39 is the median of the triangular distribution defined by the minimum \( a = 1,000 \), the maximum \( b = 2,000 \) and the mode \( c = 1,300 \).

This example is purely didactic, for a first demonstration. In the expert judgment practice, the main interest is in the repeated use of the inverse cumulative distribution function. With its help we obtain random draws from the triangles of the various experts on the way to an aggregate estimate of the mean (or median) and of some meaningful confidence interval.
ANNEX B

Common biases and remedies in elicitation

Biases and their influence on expert-judgment
All of us are inclined to interpret situations in biased ways, often based on our cultural norms and beliefs. Even the most well-trained intelligence brain is still a human brain. Biases are normal processes designed to make decisions quickly. They are unconscious, automatic and non-controllable and there is no magical solution to overcome these reflexes. Expert-judgement should be based on a systematic consideration of all relevant information and the effects of the biases should be mitigated. More than 280 biases have been identified by researchers, among which three types are particularly relevant to elicitation of expert judgement: selection, social and process biases. This chapter introduces some of the main biases that can influence expert-judgement for each category. Being aware of specific biases, and how they work in practice is the first step to mitigating these biases and a discussion of main biases should be part of any preparation for expert judgement. In addition, this note outlines some structured analytical techniques that can be used by those involved in data collection to mitigate their effects.

The process of selection the information on which experts base their judgements is vulnerable to selection biases, particularly:

- **Absence of evidence**: A failure to consider the degree of completeness of available evidence and not addressing the impact of the absence of information on analytic conclusions.
  
  Example: During the Ebola crisis, no nutrition related problem was reported. It is then tempting to conclude that no nutrition need and support existed. However, the absence of nutritional status information was a result of the “no-touch” policy which prevented screenings and therefore the reporting of information. The absence of information in this case did not indicate the absence of a problem, but the difficulty of getting the information about a potential issue.

- **Anchoring effect**: Relying too heavily on one piece of information, usually the first piece of information found, when making decisions.
  
  Example: When assessing how many people are food insecure in a given area, a working hypothesis of 100,000 people is suggested to the analyst in charge. This initial estimate is going to frame how the analyst is thinking about the situation. It would now be harder for him to sufficiently adjust up or down from this value and go up to 1,000,000 for example. 100,000 is now his anchor. When presented with an estimation task, instead of asking what is the most probable value rather ask first about the extremes, the higher and lower bound and readjust the value from there.
• **Confirmation bias:** Only seeking information that confirms our initial decisions, hypothesis, judgements or conclusions ignoring contradictory information. We tend to listen only to information that confirms our preconceptions.
  Example: The Ebola epidemic in West Africa was initially approached from a sole medical and epidemiological perspective. Only Ebola cases were seen as the priority. The initial assessment of the situation did not provide a good comprehensive picture of humanitarian impacts and requirements. It provided only a fragmented picture of the needs and risks and left organisations to neglect beneficiary groups at the beginning. This slow and inadequate perception of the crisis produced “a new crisis within the Ebola crisis”, with major non-Ebola related health, food, livelihood and education needs unmet.

• **Evidence acceptance bias:** Accepting data as true and focus more on the coherence of the story than the reliability of the underlying data.
  Example: I have multiple local and biased sources indicating protection needs: few mention recent increase in displacement, others mention reports of gender-based violence and some indicates tensions between local armed groups. The story seems to hold up: because of renewed tensions between local armed groups, the population started fleeing the area and was being targeted by the armed group. I will accept the story as it seems to make sense. However, if I start to have a closer look at the data, I would realise that the recent increase dates prior the renewed tensions between armed groups. If I dig a bit more, I might realise that no baseline data was available before on displacement so the “increase” mentioned is based on the intuition of the author of the report and not on credible data.

• **Satisficing bias or premature closure:** Selecting the first finding or conclusion that appears “good enough.” Prematurely stopping the search for a cause when a seemingly satisfactory answer is found before sufficient information can be collected and proper analysis can be performed.
  Example: In a given area, a report mentions a drop of students attending schools. After some verification, the area happens to be a territory contested by different armed groups. Now that I found a logical explanation - students are not safely able to reach schools due to the conflict and insecurity – I will stop my research and accept this explanation. However, if I continue my research, I might found out that in fact insecurity is not the main factor preventing them from going to school but actually it is the increase in transport costs. If I prematurely stop my research after the first possible and rational explanation, I might miss the actual explanation.
To mitigate these biases, triangulate the information with other sources: – are there
details being left out by one source? Assess the credibility of the evidence:

- Evaluate how accurate and precise the information is.
- Check for strong corroboration and consistency with other sources.
- Look for outliers
- Identify the key themes indicated in the evidence and assess the weight of evidence
  supporting specific conclusions.
- Consider if the explanation is plausible given the context
- Re-examine previously dismissed information or evidence.
- Consider whether ambiguous information has been interpreted and caveated
  properly.
- Indicate a level of confidence in references.

Also check if your analysis is based on sufficient information.

- Assess what information is missing and also how necessary it is to get this
  information.
- Compare how much time, effort, resources it will take to get or have access to this
  information.
- Ask yourself if you can use lessons learned or historical analogy to fill this gap.

**Group biases**

When expert-judgment is done in a group setting, several specific biases come into play.
Social/group biases are a result of our interactions with other people. The way we are
processing and analysing information depends on our relations with the persons who
provided us with information or hypotheses.

- **False consensus**: Overestimating the degree to which others agree with each
  other and usually assume that silence means agreement. **Groupthink**: Choosing
  the option that the majority of the group agrees with or ignoring conflicts within
  the group due to a desire for consensus. Belonging to the group becomes of
  greater importance than expressing individual disagreements. Members
  therefore avoid going against the flow of the discussion and do not examine
  thoroughly alternative hypothesis.

Example: The absence of negative reactions on the findings of an assessment
for example does not always mean that every member of the team agrees with
the findings. Some might be afraid of the consequences of speaking up, some
might feel they are not legitimate enough to express their disagreement. It is
easier to comfort our opinion by not seeking explicitly feedbacks.

When using the Delphi method, this bias can have a significant influence for
example. Woudenberg wrote that “a Delphi is extremely efficient in obtaining
consensus, but this consensus is not based on genuine agreement; rather, it is the result of strong group pressure to conformity” (Morgan 2013).

- **Halo effect**: Accepting or rejecting everything another group member says because the analyst likes/dislikes the person.
  Example: Affinity plays a bigger role in our analysis than we think. I will have a tendency to trust what my dear friend and colleague said rather than what my competitive and unfriendly colleague have to say about the same situation.

- **Institutional bias**: Interpreting information in line with the interests of a certain organisation.
  Example: A World Food Program employee will have a tendency to analyse information through the lens of food security and livelihood for example while a World Health Organisation worker will pay more attention to diseases and health infrastructures.

To mitigate these biases, actively seek alternative theories.

**Competing Hypothesis**: After identifying all reasonable alternative hypotheses, develop a matrix of hypotheses and input the evidence for each hypothesis to examine the weight of evidence. Compare hypotheses to each other rather than evaluating the plausibility of each hypothesis in turn. The best hypothesis is not the one with the most evidence in favour, but the one with the least evidence against.

- Brainstorm to identify all possible hypotheses.
- List all significant evidence/arguments relevant to the hypotheses.
- Prepare a matrix with hypotheses on top and each piece of evidence on the side. Determine whether each piece of evidence is consistent, inconsistent or not applicable to each hypothesis.
- Refine the matrix and reconsider the hypotheses.
- Focus on disproving hypotheses rather than proving one. Identify and assess the evidence consistent with each hypothesis to see which explanations are strongest.
- Ask what evidence is missing but would be expected for a given hypothesis to be true.
- Establish the relative likelihood for hypotheses and report all conclusions.

**Process biases**
Process bias is our tendency to process information based on cognitive factors rather than evidence. When we process information, we often display inherent thinking errors. They prevent an analyst from accurately understanding reality even when all the needed data and evidence are in his hand.
Appendices - Common biases and remedies in elicitation

- **Clustering illusion:** Overestimating the value of perceived patterns in random data. The human brain excels at finding patterns and relationships, but tends to overgeneralise. We usually confuse correlation for causation. While the two might be correlated, meaning they appear to follow the same path, they do not cause each other.

  Example: During World War II, the German military regularly bombed London. Some areas of neighbourhoods in London were hit more often than others, triggering some people to move out from the worst affected areas. Consequently, the relatively untouched areas were suspected to be home to those sympathetic to the “enemy”. However, shortly after war, British statistician R. D. Clarke analysed 537 impacts and found that there was no consistent pattern that would confirm an intention to target more specifically an area than another one; the bombs which hit London were randomly dropped (Clarke, 1946).

- **Framing:** Being influenced in our decisions by how a situation has been presented.

  Example: During the Ebola crisis, one of the coordination working groups was named Dead Bodies Management. This framed the vision of the issue in very simple terms: there is a need to dispose of the bodies to avoid further infection. This label, which neglected the important social and religious component of death in the culture of the population affected, fuelled discontent, anger and frustration towards the international community and is likely to have reduced the willingness of the population to interact with the group. This working group was renamed Safe and Dignified Burials to acknowledge and take into consideration the symbolic of death and not only the management of corpses.

- **Overconfidence:** Being too confident about our abilities.

  Example: If asked to forecast oil prices in fifteen years’ time, an economics professor will probably give an estimate as wide of the mark as an individual not familiar with the issue. However, the professor will offer her forecast with certitude. The most reliable forecasts tended to come from those who avoided black-and-white predictions and were willing to admit their limitations.

- **Selective attention/perception:** Allowing our expectations to influence how we perceive the world.

  Example: If your supervisor asks you to investigate the possibility of sanitation needs in a given context, you might come across other needs that your organisation could as well address. However, since your attention will be focused on the research of one specific needs, you will have a tendency to disregard other information perceived as irrelevant.
To mitigate these biases, examine carefully the problem using different angles and perspectives.

- **Six hats method:** Edward de Bono’s method is a parallel thinking process that helps analysts overcoming their assumption, biases and heuristics. Members of a team are assigned with a “role” to play, a hat to wear. They can more easily examine a hypothesis from different angles: neutral, emotional, creative, optimist and pessimist angles. By making the way the information is processed obvious to everyone, members of a team can acknowledge the limitations and advantages of each of the roles.

**Role of the different hats:**
- The person wearing the **blue hat** is the lead of the roleplay, he/she is the facilitator of the exercise.
- The person wearing the **white hat** is neutral. He/she expresses facts, only facts. His or her points are simple, short and informative.
- The person wearing the **red hat** is using his/her emotions, intuitions and gut feelings to approach the situation.
- The person wearing the **green hat** is looking for ideas outside the box, alternative options.
- The person wearing the **yellow hat** is looking at the different options with a positive attitude, underlying the advantages and benefits of an idea.
- The person wearing the **black hat** is criticising the different options, emphasising the risks and dangers. He/she is playing the devil’s advocate.

**Sources:**
- Daniel Kahneman, *Thinking Fast and Slow* (2011)
- M. Granger Morgan, *Use (and abuse) of expert elicitation in support of decision making for public policy* (2014)
Appendices - Common biases and remedies in elicitation

Case study – ACAPS exercise on biases
ACAPS designed a training game to raise experts’ awareness of the influence of cognitive biases on decision-making.

The participants (between 4 to 6 people per group) impersonate humanitarian experts asked to look at the current humanitarian situation in Libya and agree with their colleagues on the three most important issues in the country. The discussion is influenced by overt and concealed biases of the participants, who had previously received a ‘bias card’ asking them to apply this particular bias to the discussion. Everyone participates in the discussion in line with their assigned bias. Other group members try to guess which bias is being exercised. If a participant’s bias has been recognised, the player is out of the discussion.

Figure 36: Sample bias card

<table>
<thead>
<tr>
<th>CONFIRMATION BIAS</th>
<th>HALO EFFECT</th>
<th>STEROTYPE BIAS</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="#" alt="Confirmation Bias Icon" /></td>
<td><img src="#" alt="Halo Effect Icon" /></td>
<td><img src="#" alt="Sterotype Bias Icon" /></td>
</tr>
<tr>
<td>You only seek information that is consistent with your lead hypothesis, judgment, or conclusion. For example: I do not think that piece of information is relevant in this context / I do not believe what is said, as that contradicts everything we know.</td>
<td>You agree/disagree with the other team members depending on your affinities with them. For example: Since I like/dislike you, I agree/disagree with your assessment of the situation.</td>
<td>You expect a group to have certain qualities. For example: We all know … are usually at greater need of humanitarian assistance.</td>
</tr>
</tbody>
</table>

Lessons learned
This game, by making the way the information is processed under the influence of specific biases obvious to everyone, enables participants to acknowledge the potential damaging effects of biases on analysis and decision-making. Participants’ willingness to examine their own possible biases is an important step in understanding their impact and can propel them to act to mitigate their effects. It will not be possible to avoid the automatic bias response, but through awareness, it will be possible to consciously rectify the consequences.

This game has been tested on about 20 ACAPS Junior Humanitarian Analysts and 30 more experienced analysts from different UN organisations and International NGOs. The impact of this game has been significant on Junior Analysts, who frequently
referenced cognitive biases in following analysis meetings and discussions and trying to apply structured analytical techniques to mitigate their effects. With more experienced analysts, results have been mixed: some more senior staff are less enthusiastic in role-playing and might consider themselves already aware of their own biases, limiting the beneficial impact of this game.
References


ACU (2013). Joint Rapid Assessment of Northern Syria. Final Report [17 February 2013], Assistance Coordination Unit (ACU), supported by ECHO, DFID and OFDA.


References


Benini, A. (2015). Additional notes concerning estimates of sub-district populations in Syria, August 2015. Ghaiziantep, Turkey, UNOCHA.


References


References


References


References


References


McGrayne, S. B. (2011). The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy, Yale University Press.


References


References


References


References


The Use of Expert Judgment in Humanitarian Analysis – Theory, Methods, Applications.