August 2016

Technical Note

SEVERITY MEASURES

IN HUMANITARIAN NEEDS ASSESSEMENTS

Purpose, Measurement, Integration

Aldo Benini

Contents

Suggested citation:

Benini, Aldo (2016). Severity measures in humanitarian needs assessments - Purpose, measurement, integration. Technical note [8 August 2016]. Geneva, Assessment Capacities Project (ACAPS).

Cover photo:

Tornado forming over Lake Tanganyika. Kalemie, DRC, 2005. © Aldo Benini.

Tables

Figures

Acknowledgment

For essential help with this note, I am indebted to:

Jorge Andrés, GIS Specialist

Patrice Chataigner, Senior Analyst, ACAPS Geneva

Umar Daraz, Information Management Officer, UNICEF Amman

Agnès Dhur, Chief, Coordinated Assessment Support Section, UNOCHA Geneva

Boris Aristín González, Regional Needs and Monitoring Coordinator, iMMAP Amman

Anthony Liew, formerly UNOCHA Kathmandu

Roberto Saltori, Program Coordinator WASH, UNICEF Amman

However, all errors in this note, formal or substantive, are mine.

Aldo Benini
Consultant

8 August 2016

Summary

Severity is a key parameter in humanitarian decision making. “Severe” is part of ordinary language; “severity” more institutional. We make absolute statements (“The patient is in a severe condition”) as well as comparative ones (“The townspeople are even more severely impacted than the farmers”). The bases of such judgments are not always clear. By contrast, needs assessments in humanitarian action strive to define severity measures that are transparent, tightly related to needs concepts, and fit to support valid comparisons.

The function of severity measures is to substantiate priorities that, together with parameters like access and cost, guide decisions on the humanitarian response. Severity measures condense, in one number or one verbal scale, elements that influence judgments on priority – elements that are conceptually different, or arrive from separate information sources. Such constructs have been around for several decades, at first in public health. The humanitarian community developed its own gradually, such as in famine early warning and in the UN Food and Agriculture Organization’s Integrated Phase Classification.

For the past several years, ACAPS has assisted the development and application of severity measures. The assistance was provided in the shape of analytic tools, personnel on mission as well as remote statistical support. This note builds on key lessons from past needs assessments and recommends good practices for how to measure severity for future ones. Where we recommend more involved procedures for special situations, we refer the reader to published notes that demonstrate them in the requisite detail.

Two types of measures

Severity measures in the humanitarian domain broadly fall into two categories. Measures directly related to humanitarian sectors (food security, WASH, protection, etc.) for the most part come in the form of rating scales and persons-in-need estimates. Measures not defined in terms of sectors result from the combination of indicators that cover several dimensions of the crisis. Essential dimensions include vulnerability, intensity, exposure. Vulnerability is the degree to which an affected unit lets a given event type cause harm. Intensity is the strength, at a given time and location, of the harmful event. Exposure is scope and scale of affected units, expressed chiefly as population or area.

The distinction between the two categories of severity measures is not absolute. Some elements (e.g., populations) and some challenges (e.g., ordinal variables) are similar. Yet the analysis strategies on the way to summarizing such measures differ. Persons in need lend themselves to aggregation across sectors and hence to higher administrative levels with easy formulas; aggregating the ordinal severity ratings, however, is frequently done through statistical operations that are illegal and meaningless. In this note, we propose a new, if untested method for better results - a version of the so-called "ridit" transformation. Conversely, multi-dimensional indices use well-established methods to compress indicators into one compact measure. They enlighten us about the severity; yet the estimates of persons in need that response planners require cannot be obtained without additional survey information.

[Sidebar:] Needs, risks and severity

The concept of human need is both simple and complex at the same time. The everyday language of need – "I am thirsty; I need to drink" – is straightforward, and so is our knowledge of many of the mechanisms related to needs – we know that persons deprived of food for a long time will die although we have not made this experience ourselves.

But people have multiple needs; these defy exhaustive enumeration. A generic definition of need therefore is not obvious. It would have to relate both to the individual and to society. One could venture to say, for example, that a need is a disposition of the individual that, if met, gratifies him as well as continues the fabric of society. If the need is left unmet for significant time, it impacts the individual (mostly negatively) and may alter the fabric of society.

In the context of humanitarian crises, we want to accentuate the aspect of deprivation. We add two criteria:

- We recognize needs that have gone unmet for significant periods of time, to a significant degree, in a significant number of individuals.
- The unmet needs threaten core values including life, health and dignity of the individual and/or the institutional fabric, collective survival and long-term prosperity of the society.

This understanding poses a number of problems for the measurement of unmet needs. We emphasize two:

- **Events and risks:** We observe a crisis through what actually happened – people are no longer where they used to be (displacement), floods destroyed the crops (poor harvest), falling buildings crushed people (earthquake deaths). The ensuing threats to the living and to society, however, appear as *risks*. Risk is a possibilistic concept – more people may die over the next days, weeks, months – or not. The gap between indicators of the past – anything measured is, by definition, a thing of the past – and the likely consequences in the future (the risks) has to be bridged in the present. We build these bridges through models – the subject of this note. Models construct unmet needs as shortfalls against normative references (e.g., the SPHERE standards) or causally associated with harmful consequences. Implicitly or explicitly, both kinds transform observed events into predictions.

- **Scales and sectors:** The threats from unmet needs translate into real consequences at very different time and social scales. Thus, the absence of safety and security may result in the immediate death of the individual as soon as the threat of disaster or violence materializes. Lingering food scarcity may elevate mortality slowly and may cause a variety of health impacts noticeable only in the medium term – impacts that are unequally distributed across society. The disrupted education of an entire generation will hamper the road to recovery and prosperity for communities and nations, for decades to come. Institutionally, the various types of unmet needs are observed, analyzed and catered to by different sectors of the humanitarian system. Because of the different scales, inter-sectoral measures of the severity of unmet needs are challenging. Measures from the protection sector (immediate threats), from sectors responding to basic material needs (shelter, food security, etc.) and from those concerned with the longer-term societal functioning (education, early recovery) are difficult to combine in one overarching severity index.

We look at the first of these two challenges more optimistically than at the second. Both needs and risks are dispositional concepts. Social science has learned to work with both manifest indicators and latent constructs; recognized methodologies of index formation take care of a good deal of these difficulties. By contrast, the tradition of combining measures of unmet needs across humanitarian sectors is much younger, lacking firm technical guidelines. For example, the Syria 2016 Humanitarian Needs Overview (HNO) made a daring attempt to compute an "overall severity" rating for ten sectors for each of 270 sub-districts. The validity of such a broad measure has not yet been established in a recognized methodological and institutional consensus.

Process and measurement

For both sector-related and other severity measures, design, data collection and analysis are ultimately guided by their process and measurement models. The process model defines the basic variables of the process that affect units of interest (social groups, local communities) to certain degrees of severity; it also postulates the relationships that connect the variables. The measurement model defines how each of the variables is to be measured; most frequently this happens through algorithms that combine lower-order variables. One of the strong recommendations that grew from insight into various severity measures is that needs assessment designers should state both models – process and measurement – explicitly, in diagrams and formulas.

For sector-defined measures, the process determining the level of severity is the accumulation of destructive forces. They redistribute the population over the range from fully met need to lethal deprivation. Measurement, however, is tricky, with distributions scantily described by two measures per sector – the number of persons in need, and a qualitative severity rating on an ordinal scale. Aggregation using both measures simultaneously has been elusive, both across sectors in the locality and for a given sector nationally. In a novel approach, we propose to transform ordinal ratings to a ratio-level scale, either expert-defined or data-driven. This does not in itself ensure coherence between severity ratings and persons-in-need estimates. Little improvement will be achieved until the sectors divide area populations into estimates stratified on several grades of need. At least for the sectors catering to basic material needs, the grades can be defined by survival chances. A coordinating agency ensures that the grades are comparable across sectors. The following table suggests a notional format to fill in the graded estimates for a sector and area. At each grade, it elicits a most plausible figure as well as the minimum and maximum below/above which the sector coordinators will not go. Cross-sectoral aggregation would rely, at each grade, on multiple random draws from the plausible ranges and on formulas that an information manager can perfectly run in an Excel spreadsheet.

Table 1: Severity estimation through graded persons-in-need estimates

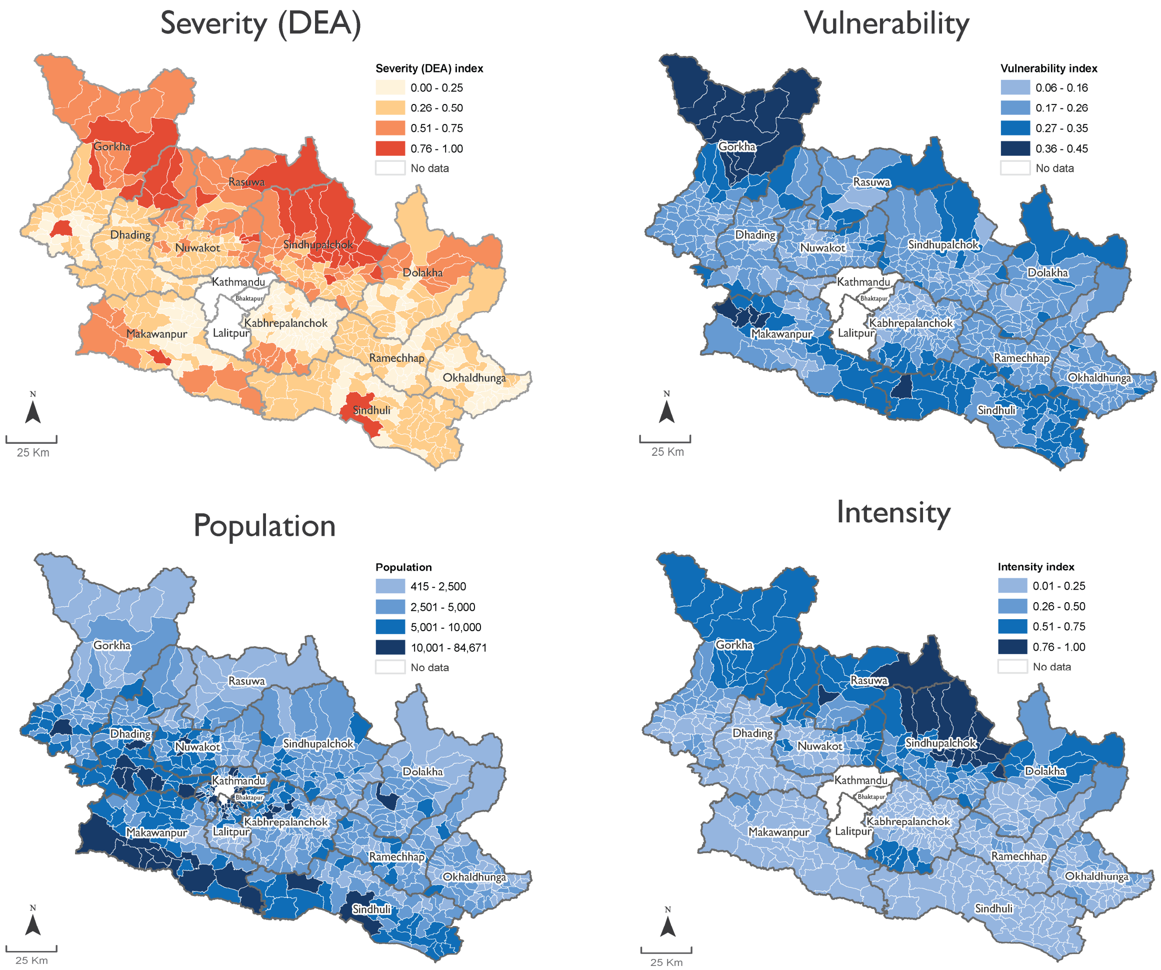
Levels of need in the _______ sector	Population and persons in need	What is the evidence for your estimate?
The estimated **total** population currently in ___________ Area	100,000 (90,000 – 120,000)	
Those with **minor to major** problems regarding the needs in this sector [will survive without assistance, but longer term damage likely]	60,000 (50,000 – 70,000)	
Those with **severe** problems [will survive if assistance given within one month]	30,000 (25,000 – 35,000)	
Those with **critical** problems [will survive if assistance given within a week]	7,000 (5,000 – 8,500)	
Those with **catastrophic** problems [most will die even if assistance is available today]	3,000 (2,000 – 4,000)	

For severity measures based on indicators that are not defined in sectoral terms, process models are primarily distinguished by the way they define the interaction among vulnerability, intensity and exposure. If these compound each other, we favor multiplicative models. If they are seen as substituting for each other, other models are appropriate. These too have been discussed and demonstrated in previous notes. On the measurement side, the integration of multiple indicators into one sub-index per basic dimension can be challenging; we present an algorithm that minimizes redundancy (i.e., maximizes the expression of the diversity).

Severity measurement and response planning

The final severity index and its distribution will, more often than not, look opaque to all but those few who constructed and computed the index and are versed in its technical arcana and in the strengths and weaknesses of the data. The consumers, particularly the response planners, may not find a straightforward interpretation. For example, the index created after the second earthquake in Nepal in 2015 was shared in population-weighted and unweighted versions, in maps that indicated diametrically opposed priority areas. Response planners may want to see not only the final severity scores, but also how these are composed. Maps and tables of assessed units should present severity scores alongside their major components. For Nepal, we have redone the analysis in great detail (pages 59 - 67); this map here throws into sharp relief clusters of more severely affected communities, together with the differential drivers of severity. It is obvious that the dark red clusters alongside the northern border are due chiefly to higher vulnerability and intensity; the smaller clusters in the southern region ae driven by greater exposure (larger populations). Such differences are important in translating severity into locally adapted, technically and socially meaningful action.

Figure 1: Nepal - Area affected by the earthquakes in 2015 - Severity and its components



Note: Calculations by the author, using indicator data published by UNOCHA in 2015. The severity index was computed under a Data Envelopment Analysis (DEA) model, with data-driven weights.

Structure of the note

An array of variegated tools is thus available for the quantitative expression of the severity of unmet needs. We review these summarily in the main part of this note. We proceed as follows: In a historic flashback, we enumerate key findings from subsequent studies. We discuss challenges and solutions for sector-defined and other severity measures. We illustrate the former types with examples from the UNOCHA-led Humanitarian Needs Overview for Syria in 2016 and the latter with severity indices produced in the response to the two earthquakes that struck Nepal in spring 2015. We conclude with an outlook on future work (pages 67 - 69) and with recommendations (pages 69 - 71) as to which methods to prefer in what types of situations.

In all that we do not advocate particular selections of measures over others – the choices will always depend on purpose, time and resources as well as on available data and accessible sources. Rather, we emphasize an orientation towards sound principles, notably the need to work with explicit models, with methods tried and tested in other fields and, at the same time, with the courage to innovate from within the humanitarian community. Progress in severity measurement began a hundred years ago. Both as an operational need and as a research program, under this or any other names, it will continue for the foreseeable future.

Introduction

The language of severity

Severity is an important parameter in humanitarian decision making. "Severity" may be about different aspects of a crisis – risk, impact, unmet needs – each with its associated vocabulary. In the needs assessment perspective, severity is about shortfalls, hence about their observed or likely consequences, and finally about necessary and possible action to remedy or forestall them. Communications on severity focus attention on levels of unmet need, but they do so frequently in the company of other important parameters, notably those of the response, such as cost, access, speed. Some of these may themselves be described in terms of severity – think of "severely limited access."

In this note, "severity" and "severity measurement" chiefly characterize levels of unmet need (although there are situations where needs can be inferred only from observed impacts or calculated risks). These concepts become fully relevant when decision makers are faced with alternatives, and beliefs about severity enter as considerations in ordering priorities. Situations, social groups, institutional sectors may be described as more or less severely affected, in vague general terms, or with regards to specific needs. Some descriptions may provide a self-contained, absolute expression of severity, as in "10,000 families lost their homes". Others make sense in relative terms only, as in observations that refugees who stay with relatives in town appear better nourished than those in camps.

The variability of disasters and crises, the specialization of institutions looking after particular needs, the competition for limited resources reinforce comparative perceptions of severity. We routinely make comparisons in natural language, and nothing stops them from being extended to the humanitarian sphere. Thus it is common to describe the sufferings of children in terms more severe than those reserved for adults. For humanitarian decision making, that has not been enough. In parallel with the increasing trust that western societies have come to place in numbers (Porter 1995), with the audit revolution that branched out from a concern merely with financial probity to organizational performance, and finally with the professionalization of the humanitarian sector, severity now has to be expressed in numbers. And for this purpose it has to be measured.

Historical origins

This has not come overnight, nor have severity measures been pioneered by humanitarian organizations. Figure 2 further below locates the first stirrings of concepts of severity measurements in the 1910s.

It is a fair guess, looking at the titles and some pages of the books in which severity measures first appeared, that it was the medical and public health professions that pioneered them. As far as the Google Books service captures the Anglophone literature, "severity rating" was first used in 1916, in a report on industrial accidents[1]. The first so-called "severity scale" was formulated for eye lesions in a chemical warfare medicine

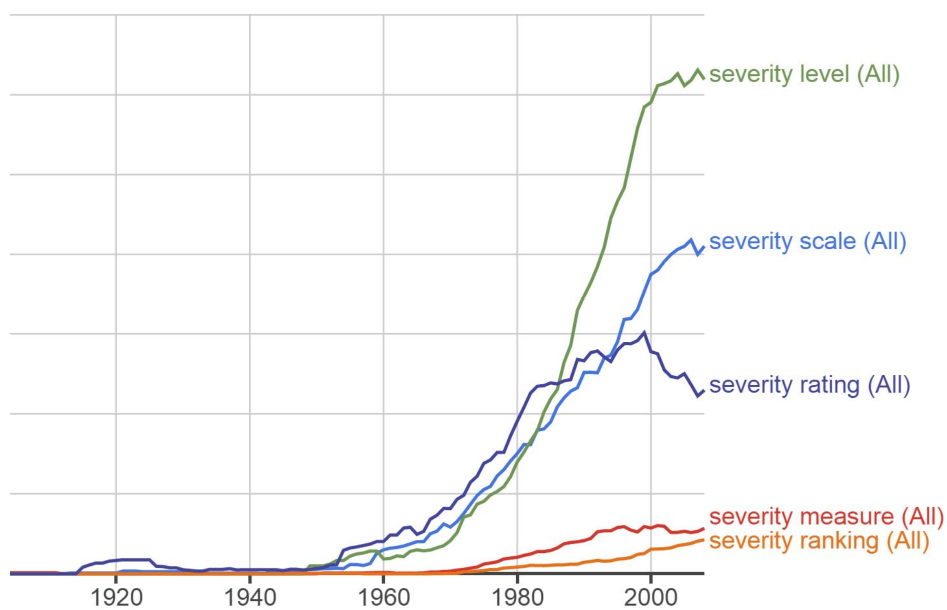
[1] Statisticians classifying hazardous occupations struggled with the appropriate measure: *"In other words, it is desirable to have the scale used as accurate as possible, but the fact that a completely accurate scale cannot be devised does not impair the value of accident severity rating"* (Bureau of Labor Statistics (US) 1916:27).

guidance in 1945[2]. A proper take-off had to wait for another 45 years. Again, it was the public health profession who blazed a trail for severity measures to move closer to the humanitarian field, with a "disaster severity scale" proposed in 1990 (De Boer 1990).

The proliferation further into the field is hard to reconstruct; we have not found seminal publications in other sectors that used the term "severity scale" demonstrably for the first time. Severity expressions certainly abound in early warning systems about famine and armed conflict; the first go back as far as British rule in India (Sen and Drèze 1999); the second multiplied after Rwanda 1994 (Adelman and Suhrke 1996, Harff and Gurr 1998). The Technical Manual of the FAO Integrated Food Security Phase Classification (IPC Global Partners 2008) can be understood as a severity scale, without definitively being called that[3]. The second edition (IPC Global Partners 2012) summarizes the FEWS Net "Food Insecurity Severity Scale" only in an appendix.

While the IPC rates food insecure areas of any definition, other initiatives use countries as their basic unit of analysis. Work by the Swedish Karolinska Institute (Eriksson, Ohlsén et al. 2015) and the European Union Joint Research Center (JRC) (De Groeve, Poljansek et al. 2015) belong here. The JRC does not use "severity scale" explicitly. The reluctance to adopt the term for the key metric indicates that these researchers feel that important methodological questions have not be settled – as their ancestors in the US Bureau of Labor Statistics did a hundred years earlier.

Figure 2: Relative frequency timelines for terms with "severity" in published books



Source: https://books.google.com/ngrams

---

2 Working with a point scheme, as often found in medical scales: *"Heavier weight is given to the more significant symptoms and therefore they receive more points in the severity scale. The sum of the points for all of the symptoms in a particular eye gives the total grade of the lesion. This may be compared with the maximum possible grade obtained from a maximal level in each symptom and thus converted to a measurement of per cent severity"* (National Research Council (U.S.) 1945:117).

3 The authors use the expression "more in line with a severity scale" five times, always in connection with renaming the "Chronically Food Insecure" (pp. 53-54), but "severity scale" appears nowhere else.

The relative frequencies of these four terms – "(All)" means that both upper and lowercase variants were counted – includes all occurrences, of which, we suppose, only a fraction were in books on humanitarian action. A detail is noteworthy: around the year 2000, the use of "severity ranking" increases steeply in comparison to "severity rating", which had had a historical lead. The growth of "ranking" methods, whatever was meant by them, would invade also humanitarian information management, with consequences that later provoked ACAPS' response (see below).

[Sidebar:] The proliferation of indices

The modern audit culture has produced a proliferation of indices in virtually all institutional sectors that collect data in indicator form. It has not spared the humanitarian community although the number of severity indices, designated as such, presumably is still small. The a.m. "Severity-scoring Model of Complex Emergency Affected Countries" by the Karolinska Institutet, Stockholm, Sweden is an example (Eriksson, Ohlsén et al. 2015).

There are many more when we widen the definition of the field. Surveying the triangle of disaster risk, vulnerability and resilience, Beccari (2016) found no fewer than 106 indices[4] created between 1995 and 2015, with a dramatic upswing after 2008. The 106 variously combined subsets of 2,298 distinct variables. More than half of the indices (81) worked at the sub-national level (62 with defined administrative units, 19 with less well defined communities), whereas twenty were set up to score countries. The point to retain is that in recent years the concerned research, assessment and response networks created and documented nearly 15 new indices per year, not to mention the unreported ones.

It is neither possible nor desirable to channel the flood of initiatives to measure concepts within that domain, which is wider than that of severity measures connected with humanitarian needs assessments. One implication for severity measurement, however, is that the prescriptive focus cannot be on the substantive selection of indicators. It must be on sound methodological principles for the combination of whatever indicators are available and related to severity (this applies to indicator-based severity measures; for persons-in-need based approaches a common approach across sectors, groups and areas is desirable – see further below).

This level of proliferating indices nevertheless should give pause. It may slow down rather than accelerate the cumulative progress of measurement methodologies and may diminish the standing of their products in the eyes of those expected to use them. Conversely, one might argue that the tradition is still so young (compared to, say, economic index construction) that only the blooming of many experiments can produce a critical mass of practitioners. Let a thousand alchemists spin gold out of all kinds of elements until a Lavoisier comes around to isolate the true humanitarian oxygen.

Other institutional domains have worked to streamline creativity into greater commonality. For example, the US public health system has struggled with fragmented systems for measuring the health of the population and the performance of care providers, made worse by growing legally-imposed measurement burdens (Blumenthal, Malphrus et al. 2015). A lack of standardization in measuring similar concepts blurred the system's ability to focus on priority issues.

Starting in the 1980s with the concept of "sentinel indicators" that capture key areas of overall improvement of public health capacity (Rutstein, Mullan et al. 1984), US public health measurement

---

[4] His terminology – "106 composite indicator methodologies" – is somewhat unfortunate. In fact, in the same article, under "methodological approaches", he distinguishes between five methods: Hierarchical methods (70 indices), Principal Component Analyses (17), Stakeholder focused methods (10), Relational analyses (5), and Novel Statistical Techniques (4).

has moved towards a small set of 15 “core measures”. These are best indicators of progress in particular areas. They reflect broader systemic changes and drive improved behavior among stakeholders. Healthy behaviors, for example, are measured by just three such indicators – body mass index, addiction death rate and teen pregnancy rate (Blumenthal et al., op.cit.: 123). De-facto, the measurement of every one of those “indicators” may be quite complex and costly, but their common focus levels the playing field plain for monitoring, evaluation and informed debate of priorities.

Of interest to our field of humanitarian severity measurement, the US health core measures were selected for their ability to meet criteria of system-wide information value. This table, from Blumenthal, op.cit.: 107, lists the criteria for individual measures as well as for the synergy of the entire set:

Figure 3: Criteria for core measures for health and health care progress, US Institute of Medicine

Criteria for Core Measure Development

Criteria for core measures	Criteria for the set
• Importance for health	• Systemic reach
• Strength of linkage to progress	• Outcomes-oriented
• Understandability of the measure	• Person meaningful
• Technical integrity	• Parsimonious
• Potential for broader system impact	• Representative
• Utility at multiple levels	• Utility at multiple levels

Some of these are directly translatable as desiderata in severity measurement (e.g., parsimony), others only partially, indirectly or not at all. But at least they deserve consideration. One has to keep in mind, though, that the US health system can impose a certain degree of information discipline in a relatively placid environment whereas the humanitarian information landscape often is quite turbulent.

Recent evolution

In a thoughtful note on the use of severity models in the response to the Nepal earthquakes, Liew (2015) invokes the continuity of the effort. Seeing a close connection with global models, he specifically anchors the breakthrough in local models in the Philippines in 2013:

> *“Severity models using composite measures are not unfamiliar in the humanitarian environment having been previously utilised for Typhoon Yolanda* [=Haiyan] *in the Philippines in late 2013 by OCHA as the Priority Focus Model, and on an on-going global basis in the INFORM model and Global Focus model. The severity model served a similar purpose in the Philippines and was used to prioritise response. At the global level, it measures risk and identifies countries with a higher likelihood for requiring humanitarian assistance.”*

Liew’s claim that the models used in Yolanda effectively guided response priorities is true as regards purpose and function. Yet, on the methodological side, a lot has changed since

Yolanda. We highlight select elements of this evolution in which ACAPS played a leading role with key points from three methodological notes.

The misuse of ranking methods

The responders to Yolanda developed not one, but four different "prioritization matrices". ACAPS analyzed them in detail, in dialogue with several authors (Benini and Chataigner 2014)[5]. Common to all matrices was the absence of a clear distinction between the process that determined the level of severity and the measurement of its components. Also, operating in a data-rich environment, the authors were faced with the challenge to standardize, weight and aggregate a fair number of indictors of various type. They addressed it by rank-transforming them, thereby obscuring the real differences between a very small number of highly impacted communities, a larger number of medium-impacted one, and the majority of slightly impacted ones.

The two main lessons that emerged from the review called for simple, but distinct process and measurement models to be drawn ahead of any composite measure construction and for the preservation of ratio- and interval-level information in such measures. The critique made an observable impact on subsequent severity models in other countries; concatenations of rankings have all but disappeared. UNOCHA's "Humanitarian Needs Comparison Tool" largely preserves measures at an interval level.

Alternatives to additive composite measures

In keeping with dominant practice in composite measures, severity models often relied on weighted and additively aggregated indicators. This index model implies that its components can substitute for each other; the weights define the substitution rates, which are constant. In process models, the drivers of severity include, at least, the intensity of the disaster, the exposure to the noxious agent (usually measured by the size of the affected population) and pre-existing conditions (such as chronic poverty levels). Whether these three substitute for, or rather compound, each other has to be determined on the merits of situation and disaster type. Even if we believe that intensity, exposure and antecedents compensate for each other, the mechanisms of substitution may be obscure. In particular, communities that score high on one of those dimensions, but low on the other two, may seem slightly impacted only, relative to others that are high on two or on all three. This additive conception may underrate the severity of the first group.

This calls for alternatives to the standard model of severity indices. If the severity dimensions compound each other, we have a case for multiplicative process models. If we believe in mutual substitution, but find constant substitution rates inappropriate, "benefit of the doubt" models may be helpful. They have been researched by the social indicator community and are suitably implemented through so-called "Data Envelopment Analysis" models. ACAPS has published a technical note on their use for severity measures (Benini 2015b).

[5] Available at http://acaps.org/img/documents/c-140527_compositemeasures_philippines.pdf .

Remaining Challenges

Challenges are, of course, changing continuously, as successes and failures invite attention to new problems and insights. Some appear to be of a more constant nature; of these we single out three that will not go away any time soon:

Process and measurement models: A generic minimal process model has not yet been formulated in a compelling manner that would act as a unifying paradigm for the majority of needs assessments. Some have cast a wide net, such as the notional model with six terms advanced for the Global Severity Index (JRC, UNOCHA et al. 2015:1):

$$\frac{\textit{Vulnerability * humanitarian outcomes * duration * threats}}{\textit{Humanitarian access * capacity to cope and respond}}$$

The multiplication and division operators suggest that all six can be measured on ratio-level scales (e.g., duration needs a defined start date; humanitarian access must not assume zero values). Also we need to clarify whether severity is to be understood primarily as impact, unmet need, or as risk. Storied crises, e.g. a persistent complex emergency overlaid by a sudden-onset natural disaster, will call for even more complex process models; the Joint Research Center (JRC) has begun investigating the conceptual challenges.

For sub-country level indices, we believe that a more parsimonious model, combining intensity, exposure and vulnerability, is better suited to finding tractable measurement models for each major term. Including exposure means that severity measures must be population-weighted[6]. For each of the three key constructs, we need to experiment with different measurement models (and will discuss some in later sections on Syria and Nepal). In particular, the aggregation of ordinal indicators remains a challenge, too often circumvented by convenient, but faulty statistical practices.

Validation: Severity measures have similarities with some clinical scales. Clinical scales too categorize indicators, weight them in some ways (often by assigning points to levels), and aggregate them (additively in most cases). These scales, however, undergo stringent validations against observed outcomes (survival, pain level, level of functioning, etc.)[7]. The current practice of severity scales hardly provides for any organized validation beyond simple face validity. Therefore, it seems desirable to do some studies that estimate the predictive power of severity measures on actual outcomes observed sometime after the onset, with appropriate controls for the intervening response. Such studies, where practical, could validate some of the theoretically and statistically significant components while dismissing others as weakly or not at all related to observed disaster impacts.

Measuring severity in **information-poor environments** poses special challenges. Indicator data will be sparse – few indicators, many missing, patently unreliable or obsolete values, definitions created for a purpose that makes the indicators irrelevant for needs assessments, etc. Anecdotal information may be plentiful, e.g. from survivor accounts, and

---

[6] UNOCHA, in June 2015, offered both population-weighted and unweighted versions side by side, with results so heavily contrasted that any response planner seeking priority areas from both would be paralyzed. While the contrast was instructive, a truly helpful multi-perspective visualization will present tables and/or four-panel maps of severity, vulnerability, exposure and intensity side-by-side. See our recommendations.

[7] Scales measuring the degree of child autism (Charman and Gotham 2013) are an example among many.

much of it may be trustworthy and compelling, if very heterogeneous. There may be few choices but to rely on the judgment of experts, where everyone with privileged knowledge of a region, population or technical domain is a potential expert.

Available resources

Since 2012, ACAPS has made several methodological notes, data entry templates and analysis demonstrations available. They can be downloaded using the links below.

Severity rating: Data management (2012)

When needs assessment teams combine data from several levels (sites, sectors and key issues), challenging data management problems arise. This note navigates through them; a companion Excel workbook template can easily be adapted for data entry.
http://acaps.org/img/documents/data-management-note--datamanagement_note-1.pdf
http://acaps.org/img/documents/severity-rating-assessment-data-management-template-severity_rating_assessment_datamanagement_template.xlsm

Severity rating: Analysis (2012)

The note details analysis strategies in assessments that identify priority sectors, key issues and severity ratings. The demonstration workbook uses data from Yemen.
http://acaps.org/img/documents/analysis-note-analysis_note-2.pdf
http://acaps.org/img/documents/severity-rating-yemen-data-analyses-demo-severity_rating_yemendata_analyses_demo.xlsx

Severity and priority – Their measurement in rapid needs assessments (2013)

The note extensively discusses the measurement of severity and priority, the analysis of such data, the differences between ratings and rankings, and the so-called "Borda count" as an ordinal-level measure of priority. It illustrates different data architectures with examples from Yemen and Syria.
http://acaps.org/img/documents/s-severity-and-priority.pdf

Composite measures of local disaster impact - Lessons from Typhoon Yolanda, Philippines (2014)

In an extensive review of four "priority matrices" built by responders to this disaster, we demonstrate valid alternatives to unfit ranking methods. The importance of explicit process and measurement models is discussed, with alternative process models and an algorithm for combining multiple indicators in sub-indices which in turn will determine severity.
http://acaps.org/img/documents/c-140527_compositemeasures_philippines.pdf
http://acaps.org/img/documents/c-copie-de-140527_philippines_demodatatset.xlsx

Moderate Need, Acute Need - Valid categories for humanitarian needs assessments (2015)

This note discusses the needs concepts and measurements. With data from Syria, it tests the distinction between persons in acute need, those in moderate need and those not in need.
http://acaps.org/img/documents/m-acaps-note-moderate-need-acute-need-valid-categories-for-humanitarian-needs-assessments-aldo-benini-march-2015.pdf

The use of Data Envelopment Analysis to calculate priority scores in needs assessments (2015)

DEA is an alternative method to calculate severity measures from indicators of intensity, exposure and vulnerability that relies on data-driven rather than user-defined weights. A demonstration uses data from Syria.
http://acaps.org/img/documents/t-acaps-note-the-use-of-data-envelopment-analysis-to-calculate-priority-scores-in-needs-assessments-aldo-benini-jul-2015.pdf
http://acaps.org/img/documents/d-data-envelopment-analysis---demo-files-jul-2015.zip

Models of severity

Basic choices

A first major distinction is between severity models directly related to humanitarian sectors and those combining indicators that are not systematically related to customary sector categories. The first type commonly generates two severity measures – an ordinal rating, with each level corresponding to a defined level of deprivation, and estimates of persons in need. For a given community, ratings and people in need (PiN) estimates then are aggregated over sectors.

The second type is driven by the opportunity to form a severity measure out of indicators that are not, or only very tenuously, matched to specific sectors, but are deemed good enough to fill a plausible process model. As such, they provide direct measures for, or form part of sub-indices measuring, its essential components – intensity, exposure, vulnerability – that together determine the severity[8].

What drives the choice?

It is questionable whether the distinction is as much driven by conceptual logic as these definitions suggest, or whether it reflects a choice largely imposed by organizational factors. The first model arises as a natural option among loosely coordinated partners who benefit from sectoral expertise centers. Each center sets the criteria for severity in far-reaching sectoral autonomy. If the crisis has made many of the earlier collected non-sector specific indicators obsolete (e.g., through ongoing population movements), the weakness of common-interest measures will also advocate the sectoral approach. The sectors do not need to subscribe to a common process model; such a model is at most implicit in the final inter-sectoral aggregation. On the measurement side, the common rules of social science apply, but may be compromised by particular worldviews or by necessary auxiliary assumptions in a very uneven information landscape.

The second model thrives on the availability of indicators and the ability of a coordinating unit to collect them in timely, consistent and area-covering manner and to weave them into

[8] A reviewer of an earlier draft pointed out that the reliance on these three essential components implies a predictive perspective in severity measurement. This is correct, for multiple reasons. Vulnerability is a dispositional construct; the damage is actualized after the measurement (e.g. last year's poverty rate) and well into the future; its relative level is predicted in the severity measure. Our knowledge of the noxious agent also points to the future of the objects that were measured for intensity and exposure. If most buildings collapsed when the quake struck, most will still be down in X weeks or months later; and most people will not be back in their rebuilt homes any time soon.

coherent measurement and process models. The role of explicit models is much stronger than in the first type. A final index delivers a severity measure at an interval or even ratio level. But this is at the expense of ready interpretation; translating the measure to a common-sense vocabulary of levels of unmet needs is not straightforward. Neither will the measure let us infer proportions of persons in need, at least not without supplementary information. On the upside, the metric allows comparisons between assessed communities in ways that are stronger than mere ordinal measures provide.

The sector-based model: The 2016 HNO for Syria

In this model, sector-based coordination structures define severity measures or interpret information already collected by others in terms of their sectoral knowledge.

Severity ratings

Ratings are based on severity scales. The details of their construction and practice matter a lot. An inter-sectoral coordination unit such as UNOCHA may create a loose common format. It will be limited to such elements as the levels, a brief verbal designation for each level, and some minimum dimensions that the sectors must fill in with meaningful criteria at the various levels.

We illustrate this with the format used in the 2016 Humanitarian Needs Overview (HNO) for Syria. UNOCHA named the labels on a seven-level severity scale, an earlier version of which had appeared in the Syria Multi-Sectoral Needs Assessment (MSNA) in 2014. It reduces to six levels when we dismiss the baseline level "0 – No problem" as purely theoretical:

1. Minor problem
2. Moderate problem
3. Major problem
4. Severe problem
5. Critical problem
6. Catastrophic problem

The sectors were to define their own criteria at each level. Yet these were to meaningfully cover all of four dimensions ("topics") that UNOCHA specified:

- Magnitude in terms of population number
- Coping mechanisms
- Access
- Availability

Some sectors elaborated on the topics, or replaced some with references to their own instruments. The degree to which the sector coordination units operationalized criteria in the grid strung out by topics and severity levels varied. For example:

- The health sector associated every severity level with a specific range of clinics in the sub-district that still reported delivering a minimum packing of services. If the percentage of such clinics fell below 20 percent, the sub-district, on this particular item, was deemed in "catastrophic" condition.
- By contrast, WASH defined availability in totally uninformative manner – moderate shortages of water were a sign of moderate problems, major shortages a sign of

major problems, etc. More meaningfully, the sector specified multiple symptoms of access problems at different levels of severity.

The sectors thus had far-reaching discretion, and for some aspects let their knowledgeable local partners determine which criteria were met at which level. In the appendix, we include a sample of a sectoral severity scale.

Ratings for individual sectors

The placement of an assessed unit – e.g., a sub-district in Syria – on the severity scale is tricky. Even if the information arriving from partners (and previously from the partner's key informants on the ground) were unambiguous, a unit might score lower on one criterion and higher on another in this composite scale. Often, however, different sources provide different or incomplete information on the same criterion. Some sources are more credible than others. It is unlikely that the double variance – across sources and across criteria – can be reduced by a simple algorithm. The more viable approach is to decide on one default severity level for the entire crisis area, or on several well-justified levels that differ by major region or social group. With those levels in mind, the assessor then looks at the entire severity information for each case. If it makes the case look worse than the idea of average severity, the assessor adjusts the rating upwards. Vice versa for a more favorable set of evidence. This behavior of "anchor and adjust" is common in decisions made with messy information. Further below, we will illustrate its intricacies with a sidebar on the WASH sector severity determination for the Syria 2016 Humanitarian Needs Overview (HNO).

Inter-sectoral ratings

In theory, the inter-sectoral coordinating unit should receive from this process an ordinal severity rating from each participating sector for each one of the assessed units. In practice, different sectors may omit rating swaths of units, or may assign them uniform ratings, particularly in conflict areas with opposing narratives. If the rating matrix is reasonably complete and informative, an inter-sectoral severity rating may be attempted for each assessed unit. The question is how.

Current practice seems to be to compute the geometric mean of the ratings. This is a statistical operation on ordinal variables that is plainly illegal. It seems to be popular because Excel provides a function that absorbs missing values (GEOMEAN). The results produce meaningless ranking tables.

There are two ways to do better. They depend on whether the coordinating body believes that the sectors used anchors and adjustments that "mean the same", in other words, a given value on the scale reflects similar levels of distress, deprivation and death. Conversely, the coordinators perceive that the sectors applied vastly different standards in assigning units to levels of severity:

Ratings across sectors are comparable: In this case, the transformation of the ordinal ratings to an interval scale is justifiable. This can be done arbitrarily or via the so-called "ridit" transformation. If arbitrary, the analysts determine interval-level values on an escalating scale that guarantees good discrimination at the higher end of severity. Thus, for example, the ordinal values could be mapped (1, 2, 3, 4, 5, 6) → (1, 2, 4, 7, 11, 16). The interval scale authorizes means and sums as operations. The differences between the means have no direct substantive interpretation. But they impose greater discrimination of cases with a bias to the severely affected. We introduce the “ridit” further below.

[Sidebar:] Why the geometric mean of severity ratings is nonsense

We have noticed the calculation of an inter-sectoral severity rating by means of the geometric mean is several contexts. Apart from the fact that the mean of any flavor (arithmetic, geometric, harmonic) on ordinal variables is illegal, the geometric mean does not serve the purposes even of those who believe they can flaunt such subtleties of statistical orthodoxy. We demonstrate this with a small numeric experiment. Five districts were rated, on the current scale from 1 to 6 (6 being the catastrophic extreme). District 1 was uniformly rated at level 4. As we move down to districts 2, 3, etc., each time we decrease a value on the left side by 1 and increase a value on the ride side by the same amount. The arithmetic means therefore remain constant (= 4); however, districts 3 – 5 now have been rated catastrophic in one or even two sectors.

As one easily sees in the table below, but which practitioners of the geometric mean likely ignore, is the fact that the geometric mean does not pick up signals of greater distress. Paradoxically, as we progress to catastrophically affected districts, the geometric mean *decreases*.

Replacing the ordinal values with an escalating scale that is ratio or at least interval-level, and taking the arithmetic means, we *capture the signal*. In the example, we use the same transformation as above, i.e., (1, 2, 3, 4, 5, 6) → (1, 2, 4, 7, 11, 16). But any sufficiently escalating scale should produce means that stand out when a sector reports a catastrophic level of severity.

Table 2: The use of the geometric mean to calculate inter-sectoral severity ratings

Ordinal scale: Geometric mean for the inter-sectoral rating

	Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Geometric mean
District 1	4	4	4	4	4	4.00
District 2	3	4	4	4	5	3.95
District 3	2	4	4	4	6	3.78
District 4	2	3	4	5	6	3.73
District 5	2	2	4	6	6	3.57

Escalating scale (ratio-level): Arithmetic mean

	Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Arithmetic mean
District 1	7	7	7	7	7	7.00
District 2	4	7	7	7	11	7.20
District 3	2	7	7	7	16	7.80
District 4	2	4	7	11	16	8.00
District 5	2	2	7	16	16	8.60

The other option, the *ridit* transformation, is data-driven. The resulting score should be escalated in a second transformation, again to bias the final score to those units that are severely affected in any sector. For this, we propose to take the odds of the ridit. The sidebar below explains intuition and mechanics.

Ratings are not comparable: What happens if the sectors apply vastly different standards in assigning units to levels of severity? We then have a situation akin to scholastic tests in which different examinees all face the same set of problems. The problems pose different difficulties. A rating expresses how well an examinee did with a given problem. The ensemble of test results simultaneously reflects the abilities of the examinees and the difficulties of the problems. For these kinds of situations, statisticians rely on Item Response Theoretical (IRT) models (Samejima 1969, Zheng and Rabe-Hesketh 2007). Translated to humanitarian severity judgments, the assessed units are the examinees; the sectors are the problems; the sector-specific severity ratings are the solution quality. The IRT delivers a combined severity value for each assessed unit as well as a difficulty value for each sector and sector-specific rating level, both on the same interval scale. However, the practical applicability of such methods is limited by the small sample size – e.g., 270 sub-districts in Syria, as opposed to tens of thousands of students in some scholastic tests. To obtain reasonably robust estimates, the less frequent values at the extreme would have to be combined (e.g. level 6 with level 5), thus frustrating the purpose of identifying the truly most severely affected across sectors. In other words, if severity standards vary greatly across sectors, an inter-sectoral rating based on the sector ratings may be worthless.

[Sidebar:] Ridit and odds transformations of severity ratings

This technical sidebar is of interest to readers who seek statistically legitimate strategies to transform ordinal ratings to a higher measurement level. It demonstrates two transformations. The first creates a data-driven ratio-level scale; scales at this measurement level have meaningful means. The second accentuates the highest rankings – in our situation particularly "catastrophic" – and thus ensures sharper discrimination. The escalation of the highest levels may be extreme, depending on the distribution, but this is so by design – to make those units truly stand out when the indicators thus transformed is averaged with others.

In social measurement in general, and in severity ratings in particular, it is desirable to find legitimate transformations of ordinal variables to a higher measurement level. This additional information cannot arise from the mere definition of the order the underlies the initial scale. It may come from extraneous information (e.g., the known mean survival time of patients as a function of symptom severity measured on an ordinal scale); or it can be data-driven, derived from the empirical distribution of the sample over the levels. Of the former kind, validated correlates of humanitarian severity measures are not known (except perhaps the proportion of persons in need; see further below). Of the latter, the little known "ridit"-transformation is a candidate for our purposes (Bross 1958, Wu 2007).

The ridit transformation

The "ridit" is a transformation of an ordinal scale based on the observed frequencies or estimated probabilities of the categories in a sample or population. Formally, assume that the categories are numbered 1, 2, .., j , .., k – 1, k, with $\sum p_j = 1$. The ridit for category j is defined as

$$r_j = \sum_{i<j} p_i + 0.5*p_j$$

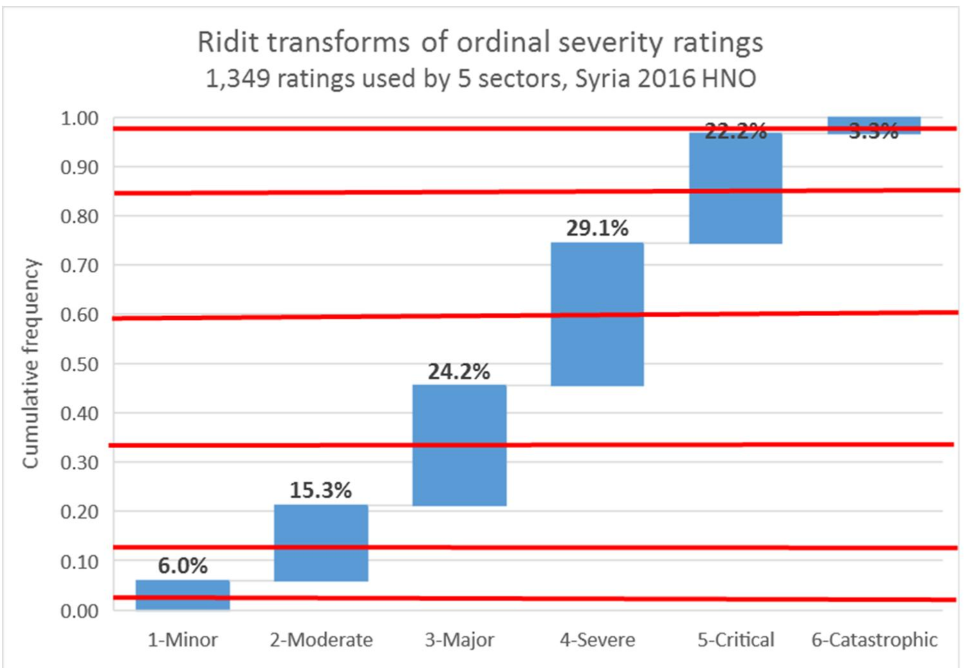
A graphic introduction is more intuitive. For the five sectors Food Security, NFI, Shelter, WASH and Early Recovery, we consider how often sector coordinators contributing to the Syria 2016 HNO used the six levels of their severity scales when rating the 270 sub-districts. All in all, they made 1,349 severity ratings, i.e., there is only one missing value.

Table 3: Severity ratings transformed to Ridits - An example

Severity level	Frequency	Percent	Cumul	CumulPc	Ridit
1-Minor	81	6.0%	81	6.0%	0.030
2-Moderate	206	15.3%	287	21.3%	0.136
3-Major	326	24.2%	613	45.4%	0.334
4-Severe	392	29.1%	1,005	74.5%	0.600
5-Critical	300	22.2%	1,305	96.7%	0.856
6-Catastrophic	44	3.3%	1,349	100.0%	0.984
Total	1,349	100.0%			

For an intuitive grasp of the ridit, we turn to the following chart. The red lines cut the frequency boxes into halves; the projections onto the y-axis (red lines) are the ridits – the cumulative frequency of all categories below, plus half the frequency of the category in point.

Figure 4: Graphical explanation of the Ridit



The ridits themselves do not help us to discriminate in favor of severely (critically, catastrophically) affected units or sectors. But they have two convenient characteristics. First, as long as the extreme categories (minor and catastrophic) are used in the ratings at least once, their ridits are > 0 and < 1. This is a statistically advantageous situation, chiefly because it avoids division by zero when we need to divide by the ridit or by (1 – ridit).

Second, the ridit is the probability that a randomly drawn unit has a severity no higher than "the midpoint" of its originating ordinal level. To give an example: The ridit for the "critical" level (level 5) in this sample is 0.856. Assuming that even within this category, all units can, in theory, be completely ordered as more or less critically needy, 0.856 is the probability that a unit belongs to any category 1 to 4 or to the less needy members of category 5, up to the median of its internal grading. Conversely, 1 - 0.856 = 0.144 is the probability of belonging to higher categories ("Catastrophic" in this example) or to the members of "critical" above its internal median.

The odds of the ridit

That is the basis for a second transformation, the "odds" of the ridit (Wikipedia 2016b). These odds are defined as

$$oddsRidit = ridit / (1 – ridit)$$

which in our example for category 5 amounts to 0.856 / (1 – 0.856) = 5.95. This is the ratio of units less severely needy than the median member of category 5 to those more severely needy. This quantity thus estimates the selectivity of extreme situations: for a given level of the severity scale, how many more units (districts, people, etc.) are in a less severe condition than "us who are stuck in this misery or even worse" – how extreme is our position compared to the relatively better off?

Table 4 gives the ridits and their odds for the observed sample of severity ratings; Figure 5 shows the simplified distributions of the twice-transformed severity measure for the five sectors.

Table 4: The odds of the ridit

Severity level	Frequency	Percent	Ridit	Odds of the ridit
1-Minor	81	6.0%	0.030	0.031
2-Moderate	206	15.3%	0.136	0.158
3-Major	326	24.2%	0.334	0.501
4-Severe	392	29.1%	0.600	1.498
5-Critical	300	22.2%	0.856	5.954
6-Catastrophic	44	3.3%	0.984	60.318
Total	1,349	100.0%		

It is obvious that the odds produce a steep escalation. In this ratio-level measure, "catastrophic" is elevated above "minor" by a factor of 60.318 / 0.031 = almost 2,000; the factor from "critical" to "catastrophic" is 10.1. These values are data-driven; the more readily the assessors grant the rating "catastrophic", the more its ridit will fall, and with it the odds. By contrast, if this rating is used rarely, ridit and odds will skyrocket. There is nothing wrong about the arithmetic; the question is whether policy makers are willing to put catastrophically affected units into such stark relief. Do they trust that the various sector coordinators assigned this level with equal restraint?

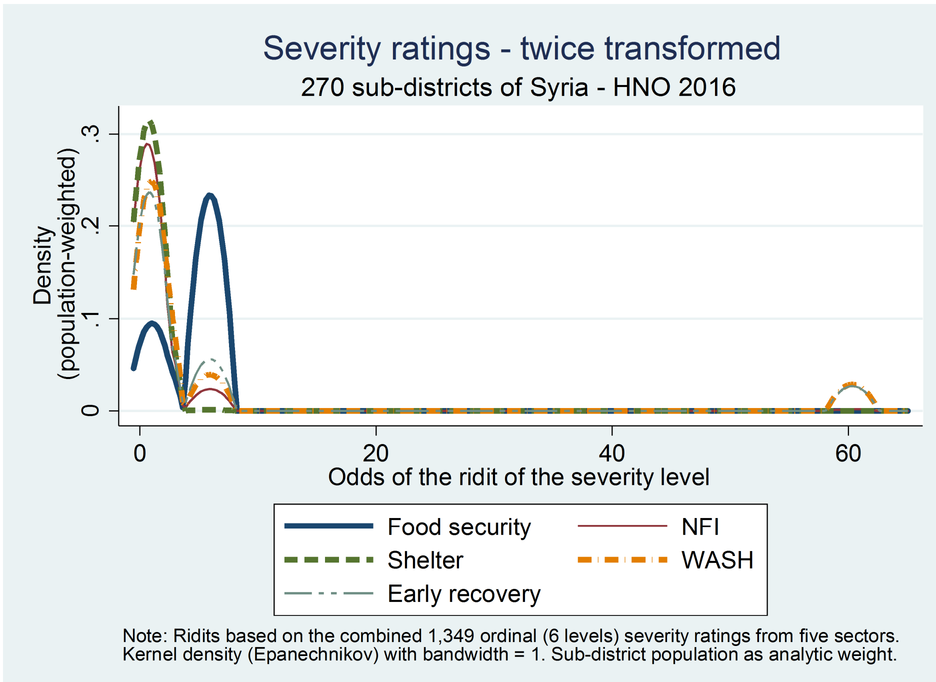
If the transformed value at the highest level is unacceptable (too high, as a result of very rare ratings at the highest level), one may consider combining the two highest levels (i.e., 5.a.-Critical/Catastrophic). Technically, this amounts to a simple recoding operation; it will replace the odds of the ridit for those two levels with one that is slightly higher than the current value of the second-highest level. Substantively, however, this backdoor escape from a frightening escalation distorts the intent of the raters who assigned the highest level.

Visualizing differences across sectors

The following graph accentuates the perception of the severity distribution by spreading out the severity values. At the same time, it simplifies perception by combining the lowest values. Peaks appear at three locations. The leftmost kind engulfs the values from 0.03 to 1.5, in other words: from "minor" to "severe". The second set of peaks, at 5.9, captures the "critical" level. Far, far out to the right is the location of "catastrophic". The heights of the peaks reflect the population-weighted frequencies.

This set-up reveals very different "distributional signatures" across the five sectors. We highlight three types with thicker lines. Shelter is limited to one peak, at the lowest values of this severity measure. WASH has as low, but not negligible middle peak. WASH also has a similarly tall right peak, indicating that sizeable populations are in a catastrophic situation (notably Jebel Saman in Aleppo, population > 1 million). Food security presents the tallest of all middle peaks; the larger part of the population is in critical condition food-wise, but none of the sub-districts was deemed catastrophic.

Figure 5: Population-weighted distribution of the transformed severity ratings, by sector



Comparing sectors

Eventually, the transformations are worthwhile only if we accept that the resulting ratio-level measure expresses severity with satisfying validity. The validity must be assumed both within and between the sectors. If so, we are entitled to compare sector means, as in this table.

Table 5: Sector means of the twice-transformed severity measure

Sector	Obs	Mean of the severity measure	
		Unweighted	Population-weighted
Food security	270	4.06	4.46
NFI	269	1.47	1.39
Shelter	270	0.70	0.80
WASH	270	3.69	6.63
Early recovery	270	9.44	6.45

Comparing areas

The sector means are column means. What about the row means - the means over all sector-wise severity values for each sub-district? The question naturally arises whether they would be equally valid – a one-stop severity measure, collapsing all sectoral differences.

What kind of policy considerations would such a combined score inform? For example, would a sub-district rated low on the unmet needs of an immediate physical kind (food security, NFI, shelter and WASH in our example), yet high on early recovery needs (a more distant consideration) deserve special attention? If so, probably in the early recovery context, and then this aspect might as well be studied with a single focus.

Averaging sectoral severity values by sub-districts therefore depends on the policy concern at hand. If we agree that for a given concern sectors A, B, C .. are relevant, the row mean of the odds-of-the-ridit severity measures is certainly more informative, and legitimate, than the often practiced geometric mean of the ordinal ratings. It gives due weight to critical and catastrophic situations.

To repeat: what is legitimate is not the row mean over any and all sectors, but only over those sectors that are of interest for the policy question at hand. "All-purpose" ranking tables are of doubtful validity.

What have we achieved?

The double transformation of severity ratings creates a ratio-level measure. Comparisons of means are thus legitimate. If the attributes "critical" and "catastrophic" are given out sparingly, the measure discriminates heavily – it ensures that units, even if they are critically and catastrophically affected in only one sector, stand out. If there is "grade inflation", units with multiple sectors at "severe" may outscore those with a single high rating and all other at "major" or lower.

The viability of this measure thus hinges on the consistent use of the ratings. In particular, if all sector coordinators share a similar understanding of what constitutes "critical", and what "catastrophic", then the transformation of the ordinal scale (severity ratings) to a ratio-level measure (odds of the ridit) promises analytic gain. Otherwise we delude ourselves with statistical alchemy, an attempt to spin fine gold out of rough straw.

Linear or multiplicative combinations of the odds of the ridits across sectors are ratio-level too; thus they can be entered into models that require this level, such as data envelopment models.

The next step then is to combine severity measures with population size, or even more informative, with persons in need.

Calculation of the ridit in Excel

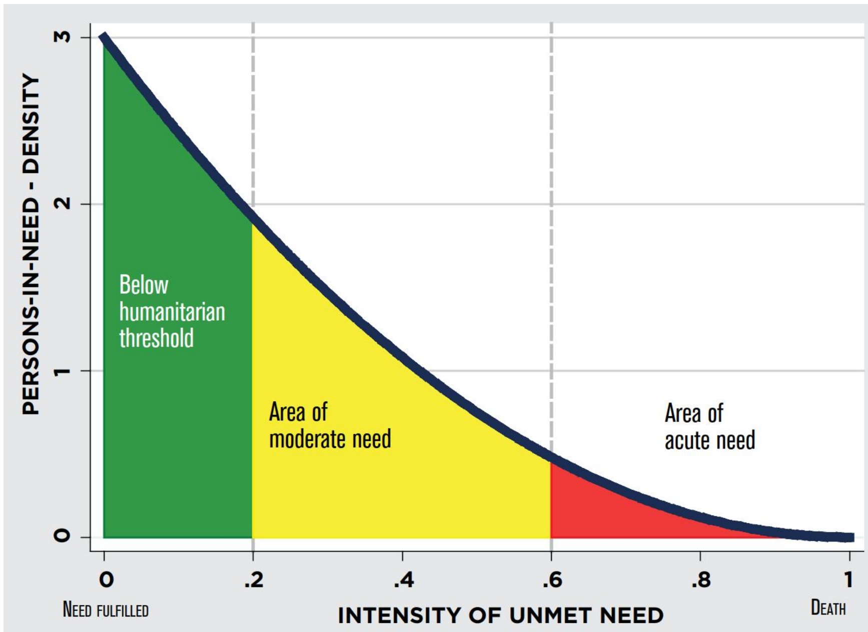
To calculate the ridit in Excel, one names a combined range for the columns holding the severity ratings of interest. In a separate sheet, the function COUNTIF is used to create the frequency of all the ratings by level. The ridit is formed at each level as the sum of the cumulative frequency up to the preceding level plus half of the frequency of the level in point. The odds of the riddit are conveniently figured in the same small table, just to the right. Back in the main table, for each sector two variables, ridit_XYZsector, and OddsRidit_XYZsector, are created. Ridit and its odds are imported via VLOOKUP.

Persons-in-need estimates

Quality and quantity

Unmet needs cannot be observed directly; they are latent variables inferred from observed behaviors. These observations inform key informant judgments, registrations for relief and other operations that generate information related to needs in one way or another. Plausibly, the intensity of unmet needs varies from "no need" to "death as a result of deprivation". The distribution of the population over this range can take variable shapes. If the intensity has a metric (e.g., the probability of death attributed to a particular unmet need), it can be represented as the distribution of a continuous variable. This chart exemplifies a possible distribution – one of many.

Figure 6: Distribution of the intensity of unmet needs - An example



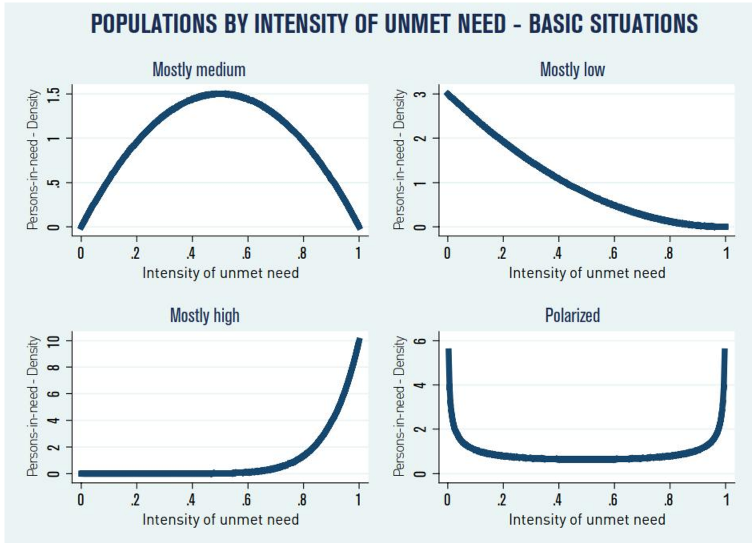
Metric intensities are rarely known – nutritional surveys come to mind as a possible exception -, but judgments about graded, i.e. categorical needs are common. We say things like "people are in acute need of safe water" or "food shortages are common, but are not life-threatening" – expressions that delineate ranges of intensity. Quantitative estimates of the population in a particular range – e.g., "40 percent of households face daily food shortages" – do not fully describe continuous needs intensities. However, if we can estimate the proportions of the population in more than two mutually exclusive ranges, we gain non-trivial insight into the qualitative shape of the distribution.

Acute needs, moderate needs

This is possible with estimates in just two grades, which we call, for want of standard terminology, "acute needs" and "moderate needs". The two proportions imply the proportion of those not in need. The three quantities, with two degrees of freedom, are enough to determine the qualitative shape of the distribution.

We expect four different basic shapes, as shown in this diagram. A fifth, the uniform distribution, is unlikely in practical life.

Figure 7: The intensity of needs - Basic shapes of its distribution

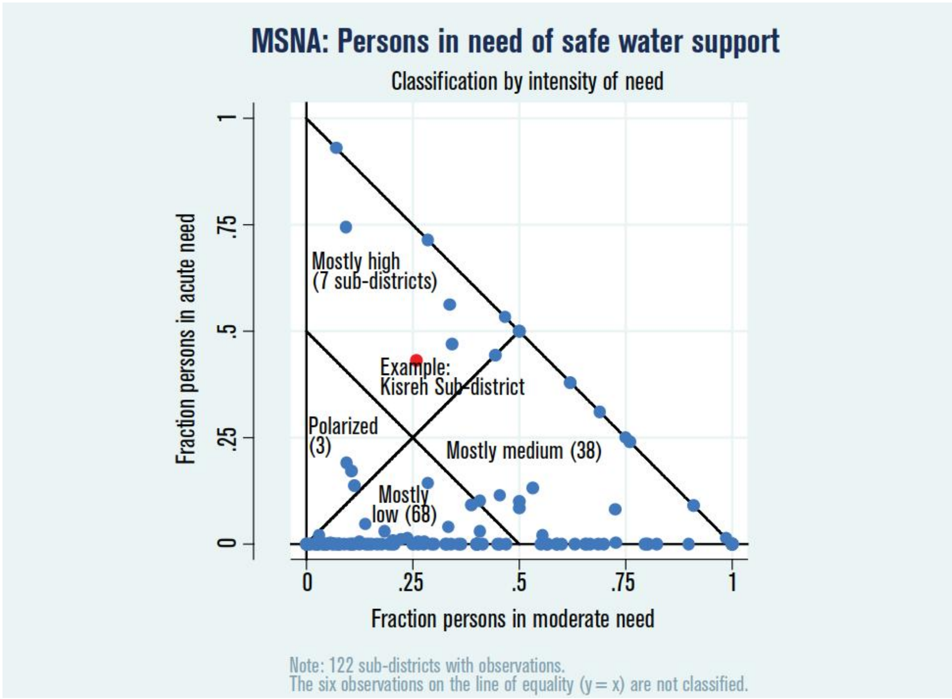


Thus, with a minimum of estimated proportions – basically two – of persons in need, we can say something important about the likely shape of the distribution of unmet needs. Statements such as "Most households in this district do not have enough food, but few are starving" use verbal quantifiers ("most", "few") (Borges and Sawyers 1974). They are inferred from incompletely observed distributions, on which we have barely enough information to characterize them in qualitative terms (Kuipers and Berleant 1988). That is, we can guess the rough shape of the distribution – such as "polarized" -, without knowing its quantitative densities. Such reasoning has the potential to become, through subsequent updates, increasingly detailed, quantitative and precise[9].

Similar distinctions have been used in food security assessments. In a few instances, they have been tried out in multi-sectoral assessments. In 2014, the Syria Multi-Sectoral Needs Assessment (MSNA) (Humanitarian Liaison Group 2014) had its enumerators elicit both moderate and acute needs estimates in five sectors. Consistent estimates of persons in need were received from key informants in 128 sub-districts. This chart illustrates the allocation to the four basic types for the WASH sector.

[9] This is what engineers and artificial intelligence researchers commonly mean by "qualitative reasoning" (Forbus 1996), distinctly from the social sciences, in which qualitative reasoning abstains from quantitative development.

Figure 8: The intensity of need for safe water - Classification of sub-districts in Syria, 2014



The chart is easier to understand with the help of an example: Kisreh Sub-district in Deir-ez-Zor Governorate (the red dot) reported a population of 116,000. The enumerator brought back estimates of 30,000 persons in moderate need of safe water support (approx. 26 percent), and of 50,000 in acute need (43 percent). This leaves out 31 percent who then were not in need of this support. This earns Kisreh the membership in the "mostly high intensity" category although 43 percent is less than half. Since the exact cutting points between the categories are not observed, the boundaries defining the four areas (black lines in the plot area) are reasonable approximations.

Subsequently, we were able to demonstrate that the separate estimates of persons in between acute and moderate need had analytical value (Benini 2015a)[10]. The distinction passed two tests. We correlated the proportions of persons in need between sectors; the correlations were stronger for persons in *acute* need than for *all* (moderate + acute) persons in need. Plus, the proportions in acute need predicted the sectoral severity rankings in addition to, and independently of, the effects of the proportions of all persons in need.

In conclusion, the concept of "acute need" is valuable. It can improve the measurement of unmet needs under conditions that rarely permit exact classification. The MSNA experience suggests that, whenever feasible, needs assessment should endeavour to estimate at least three components – persons not in need, persons in moderate need, and those in acute need. In our recommendations, we will even go to four levels.

[10] Available at http://acaps.org/img/documents/m-acaps-note-moderate-need-acute-need-valid-categories-for-humanitarian-needs-assessments-aldo-benini-march-2015.pdf .

Estimates by the sectors and their combination

The MSNA was a "survey in one hand", meaning that the enumerators sent to the sub-districts in Syria collected information on all the enumerated sectors and reported it in one questionnaire. They evaluated claims made by key informants, including claims to persons in need, from vantage points close to their sources. The claims were again evaluated in intense enumerator debriefings.

This privileged arrangement, as far as evaluation for consistency goes, is not granted to all multi-sectoral needs assessments. Sectoral information collected in the field may originate organizationally separated, flow upwards in separate channels, be processed separately by specialized sectoral coordination units, and be combined and aggregated late in the process after it has been transformed to significant degree and with distinct formats and formulas.

The sectors on their own – pros and cons

Similar inconsistencies may build up when the information on the ground was initially collected in a unified format, but was later parcelled out to be evaluated by separate sectoral channels, each proceeding with their own professional standards. Applying these standards may make the interesting parts of the information technically compliant with a sector's particular information needs. A sector coordination unit may also have specialized information that it can link to generic variables through auxiliary assumptions. These additions may strengthen the sectoral severity estimates.

On the downside, the separate treatments may render information less comparable across sectors. Worse, they may accept information that is less reliable, for lack of close data inspection in a comparative perspective early in the process. And worst of all, competition for funding may convince some sectors that they should aim for the highest credible, rather than at the most plausible, persons-in-need figures.

Bringing it all together

The question then arises what an inter-sectoral coordination body can salvage of this disparate information. What scope is left for valid severity measures, for comparison as well as for aggregation purposes? A general answer is hard to formulate because set-up and intermediate processes vary greatly. For a case study about how fragmentation aggravates person-in-needs estimates, again the Syria 2016 HNO does not disappoint.

UNOCHA engaged as many as ten sectors in producing estimates of persons in need for 270 sub-districts[11]. Both moderate and acute needs were considered, albeit in different terms. Across the sectors, the HNO estimated 13.5 million people to be "in need of humanitarian assistance". Of these, 8.7 million were "in acute need of multi-sectoral assistance." (UNOCHA 2015:6). This number is the same as the number of food-insecure people. The 8.7 million food-insecure included all the estimated 6.5 million IDPs as well as 2.2 million other food-insecure persons[12]. Why the estimate of the food-insecure was used as a proxy for those in acute need of multi-sectoral assistance is not clear. The median sub-district severity rating is higher for food security than for any other sector; this may have

[11] Syria has 272 sub-districts. Two on the Golan Heights were not covered.

[12] As explained by a member of the UNOCHA office in Amman, phone conversation, 22 June 2016.

made this sector appear particularly suitable to provide some basis for the multi-sectoral figures.

Internally, UNOCHA assumed the number of persons still remaining inside Syria to be around 16.9 million. Thus, retranslated to the three-category concept of persons in need, the HNO 2016 implies 8.7 million in acute need, 4.8 million in moderate need, and 3.4 million not in need of humanitarian assistance.

Those distinctions were made at the aggregate national level. At the local, i.e. sub-district level, the sectoral estimates worked with only two levels of humanitarian need – "in need" vs. "not in need". Nonetheless, this left the sectors with the difficult task to each provide 270 PiN estimates. In this sidebar, we present a case study of how difficult it can be for a sector coordination unit to produce such estimates. Then we discuss the challenges of comparing and aggregating across sectors, particularly when some of the sectors limit themselves to uniform, or otherwise constrained, estimates for all areas.

[Sidebar:] Severity and persons-in-need in the WASH sector

Background

In the run-up to the HNO for Syria in autumn 2015, the WASH sector, as all the other sectors, contributed its own sector-specific sub-district-based severity ratings and persons-in-need estimates. And, as most of the others, the WASH analysts, based at UNICEF Amman, depended on datasets collected outside their direct line of implementing partners. Most of the information was provided by the Whole-of-Syria-Approach (WoSA) key informant survey in July and August as well as by the ongoing Area of Origin survey. The UN Habitat Urban Profiles and the Syrian Red Crescent Society (SARC) also contributed.

In this situation, the WASH analysts found themselves working with proxy indicators, from which they deduced, with the help of simplifying assumptions, broad estimates of indicators that they subsequently combined into the two indices of interest – severity ratings and persons-in-need (PiN) estimates.

Key proxy measure

At the root of both of them lay a key distinction between high-risk and low-risk water sources, the first comprising tankered and open-body water, the second piped and bottled water. The distinction was imprecise; piped, but unchlorinated water is high-risk; some water tankers may by hygienic. Information on these subtleties was not available from the contributing surveys. WASH assumed that most households purchased tankered water, to top up on limited piped delivery. Damascus city, with wide availability of bottled water, was an exception.

The surveys had asked key informants to check predominant water sources in their respective sub-districts. The question was multiple-choice; quantities or user proportions were not elicited. The WASH analysts needed above all an estimate of the proportion of high-risk water users. They let the ratio of all checked high-risk options to all possible options among all key informants from a given sub-district proxy for this (the counts were weighted by the confidence levels assigned to informants). The mean value of the proxy was close to a WFP-VAM statistic of the proportion of households topping up with tanker water, which WASH took as a validation of its – admittedly crude – method.

Severity rating

The severity rating was significantly determined by the high-risk water user proportions, but was influenced also by other variables, to the extent that these were observed:

- The variance in the cost of tankered water in the sub-district over time (most observations were made by the Area of Origin survey) – high variance indicated irregular supplies, nudging up the severity rating
- User perception of water quality (in the end not used because of too little variability)
- Water treatment in the households – if many households were reported to resort to their own treatment, this was taken as indicating low quality, sending ratings upward.

An important feature of the WASH severity rating determination was its reliance on a starting value for average conditions prevailing across sub-districts, for which the analysts selected level 4 on the scale from 1 to 6 – designating a severe, but not critical or catastrophic problem. If the holistic evaluation of all proxy indicator values for a given sub-district suggested a better than average situation, they lowered the rating; conversely, if the information at hand suggested greater than normal deprivation, they adjusted upwards. This is an excellent illustration of anchoring behavior in a multi-criteria imprecise-information context; we will revert to this point.

The draft ratings by the analysts were run by UNICEF's partners. The agreement was very high; only for three out of 270 sub-districts did the comments motivate changes.

Severity level 4 therefore became the modal as well as median value in the WASH severity scale (134 out of 270 sub-districts), lower than food security (median = 5), and higher than shelter (3).

Persons in need

The persons-in-need estimates were produced via the estimated proportions of PiN. For each sub-district the proportion was estimated as the proportion of high-risk water users (see above) plus the proportion of IDPs (with an obvious ceiling at 100 percent). The absolute number was then simply the result of multiplying by the estimated population. The rationale was that most IDPs were likely to have to purchase tankered water.

The addition of the IDP proportion moved the WASH PiN estimates considerably upwards. The mean proportion for the 270 sub-districts is 71 percent (72 percent when population-weighted). Food security, with a higher median severity rating, produced PiN proportions with a mean of 47 percent.

Operational value

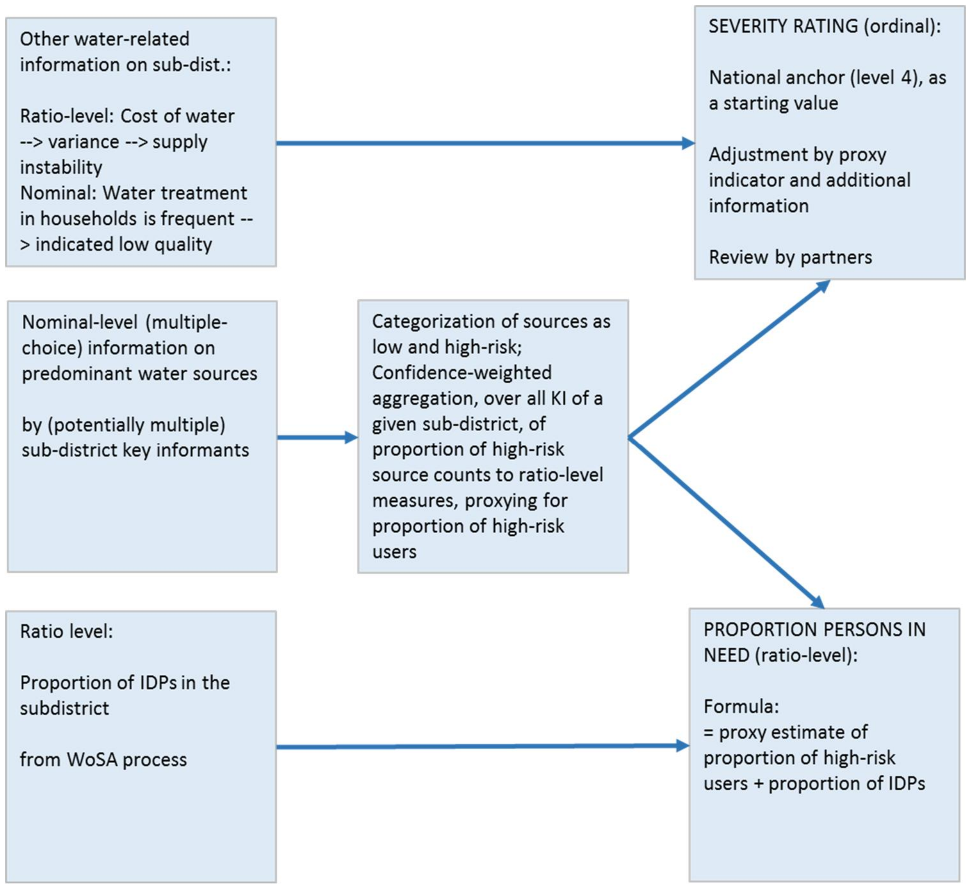
The WASH sector participated in the estimation of severity and of persons in need as a contributing partner of the HNO effort. Given the information situation, reliance on proxy indicators was the best available, and even the only possible, approach. However, severity ratings and person-in-need estimates, produced with sub-district level data, have no direct operational value for the sector. Such information, to be of value, would have to be higher resolution, collected at the household level, with instruments that truly reflect concepts of WASH professionals. Current efforts, by others, to drill down from the sub-district to the community promise an improvement over previous limitations, but do not give WASH enough voice to ensure relevant and valid measures.

Comment

The production of severity ratings and PiN estimates in the WASH sector is a fascinating illustration of information management. The sector manages the information in an environment with multiple sources, variable reliability and extraneous measurement definitions over which it has insufficient control. The dependence on a cognitive anchor – severity level 4 – and the final ratings resulting from adjustments considering various information sources is a well-established behavior in such situations (Kahneman 1982, Slovic 2000, Wikipedia 2016a). In fact, it is hard to see how the WASH analysts could have done without a normalizing anchor.

In great simplification, the process, strictly for this sector and this HNO, can be summarized in this diagram.

Figure 9: Original information, by measurement level, and transformation to indices - WASH sector



Another immediate implication from this process is that severity ratings and PiN proportions are not independent measures. They share an influential contributor (the estimate of high-risk users). This finding is not the same as saying that they are correlated – which one would expect even if they were independently constructed. Given the dependence by design, PiN proportions cannot be used to validate severity ratings.

The addition of the IDP proportions does not make the PiN estimates more reliable; it just "errs on the side of the victims", which is morally correct, but may bias the measure.

Of great concern is the difficulty in producing severity measures that are operationally meaningful to the individual sectors. The WASH analysts correctly pointed out a dilemma of information economics: Either one collects finely-grained operationally relevant information (household-level sample surveys designed by the sector) covering only part of the affected region. Alternatively, one relies on coarser estimates from higher levels of administration from the beginning (key informant judgments at the sub-district level), covering the entire region, but at the expense of operational relevance, conceptual clarity (as seen by the sector) and reliability.

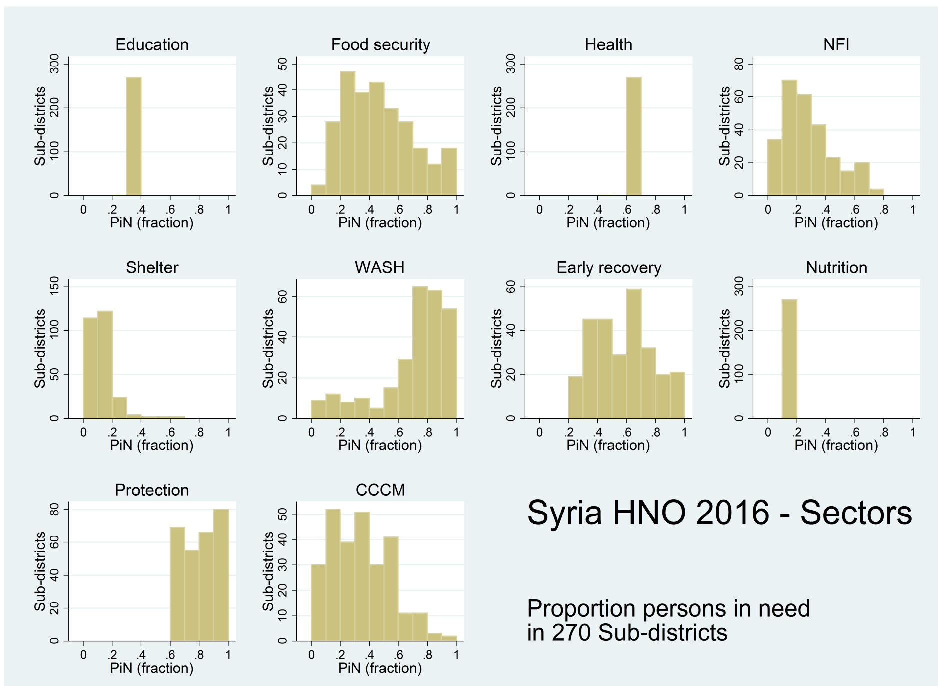
In one sentence: Both sectors and inter-sectoral coordination share a common concern in severity measures, yet with manifest tensions over usefulness and design control.

[This sidebar is based on an extensive telephone interview with UNICEF WASH coordinators in Amman in mid-March 2016, on prior email exchanges as well as on the severity scale reprinted in the appendix.]

Treatment across sectors in the face of inconsistent formats

Coming back to the contributions of the ten sectors to the Syria 2016 HNO, Figure 10 shows the sectoral distributions of PiN proportions over the sub-districts, before population-weighting.

Figure 10: Proportions of persons in need, in ten sectors - Syria 2016 HNO



Four sectors constrained their estimates. Education, Health and Nutrition found it impossible to differentiate across areas; they each assumed a constant proportion of the local populations to be in need – 34 percent for Education (based on the proportion of school age youth), 69 percent for Health (reasoning unknown), and 19 percent for Nutrition (based on vulnerable age groups). Protection is remarkable for setting a minimum, 69 percent, the same value as the constant chosen by Health. This was interpreted by some as making a political statement rather than contributing an operational measure.

The constant proportions advanced by the Education, Health and Nutrition are useful in signalling the relative acuteness of the unmet needs in these sectors. The health needs are more pressing than education and nutrition needs. Protection needs, with a median proportion of 83 percent people so affected, are even more extreme. However, the failure to differentiate across sub-districts devalues the person-in-needs estimates in the ensemble of the ten sectors. Constants are uninformative priors, and estimates truncated at a politically motivated minimum create upward bias.

The difficulty that the health sector experienced in estimating local persons in need is particularly tragic. Together with the estimates from Food Security, NFI, Shelter, WASH and Nutrition, Health sector estimates would have combined for a well-rounded measure of neediness in the sphere that generally responds to immediate material relief. Without Health, such a sub-index is imbalanced.

Similarly, Education, if local estimates were provided, and Early Recovery together might yield a useful construct for longer-term needs. Protection and CCCM might form a composite expressing immediate non-material needs.

Combining the sectoral estimates – or not?

What can be done in this situation? There are approaches to make the best of the usable information; here we discuss three, each with its pros and cons. At this point, we want to remind the reader that this is a case study; in other crisis contexts and data situations, some of what we are going to describe might not work at all, but different approaches might.

The approaches to be explored are:

1. Combined (i.e., inter-sectoral) persons-in-need estimate: The key interest is to have a figure for all the persons in need in local areas, and finally in the nation, or crisis zone.
2. Alternative severity measure: We exploit the PiN proportions as indicators to form an index of total deprivation for each area. We do without combined PiN figures.
3. Proxy indicator for persons-in-need: We abandon the PiN data altogether and substitute some demographic indicator that is not by design related to needs, such as the proportion of IDPs.

Combined persons-in-need estimate

Individuals can be in need in several sectors at the same time. The combined PiN number is the number of individuals in an area that are in acute need in at least one of the enumerated sectors, but are counted each only once. What is needed is not the simple addition, but the size of the union of sets of individuals. Unfortunately, the degree of overlap across the various sectors is not known. Technically, the size of this union is hard to estimate[13]. Moreover, measurement error – the deviation of the estimated proportion from the true proportion – is large. In this situation, for a given area or community, the *maximum* among the numbers of persons in need across sectors is a reasonable estimator for the total number of persons in need. It is reasonable because 1. in the absence of measurement error the maximum is an under-estimate (some people with acute needs will not have needs in the sector with the largest number, but in some other sectors, and thus have to be added), 2. when measurement error affects several random variables, their expected maximum

[13] "Ecological correlation" models (Robinson 1950) would be the first candidate that could potentially gauge the associations of the various sectoral needs at the individual level. Applications are more common in the health field, and almost unknown in the humanitarian (Elcheroth, Penic et al. 2013). The Syrian sub-districts, with their vastly different population sizes, make this a very unpromising procedure; yet it is not clear what other strategy would be better or easier.

exceeds the expectation without error, and this compensates for the under-estimates[14]. The degree of compensation is unknown; thus the estimate is not error-free.

Sectors with constrained PiN proportion should be excluded from this calculation. If their fixed proportion is lower than the mean of most other unconstrained sector proportions, they have little effect. If it is higher – such as for Health and Protection -, they exert significant upward bias.

The estimated total number of persons in need in Syria, using this estimator for all ten sectors, is 13.6 million. Excluding the four sectors with constrained proportions, it is 12.9 million.

Alternative severity measure

In this approach, the sector coordination units are simply agents applying a measurement instrument repeatedly to estimate the degree of deprivation. Their instruments "happen" to produce estimates of proportions of persons in need, each in their substantive domain, but this is ultimately irrelevant as long as we can believe that between them the measurement agents cover varied deprivation aspects. Thus the Food Security PiN estimate could be replaced by the Livelihoods PiN estimate if such a sector existed – the sectors are mere instances from a universe of (potentially infinite) arrangements to estimate persons in need. What matters is the common orientation towards deprivation as the degree of unmet need. The various forms of deprivation are the consequence of a common factor (the war in Syria). Because the available sector PiN estimates express a broad, but not necessarily the full spectrum of unmet needs, we seek an aggregation that minimizes overlap (maximizes diversity). This translates to the challenge of finding weights of the sectoral PiN proportions that maximize the diversity.

The "Betti-Verma double-weighting rule", developed by deprivation researchers (Betti, Cheli et al. 2005) and explained in detail by Benini and Chataigner (Benini and Chataigner 2014) achieves that. We calculate the measure using the PiN proportions estimated by the six sectors that did not constrain them. At first, we present the weights given each sector for the noted purpose.

[14] The interested reader can easily simulate this in a spreadsheet, taking row maxima of persons in need across sectors (a row represents a local area) and calculating the sum of the maxima (the number of all persons in need in the country). Assume, for the sake of the experiment, that those are true values. Then create values with error: Multiply the same persons-in-needs numbers by (0.5*RAND() + 0.75) (this expression has expected mean = 1). Take the row maxima and sum them. This sum should be higher than the first. If some areas have exceptionally large populations, the experiments may need to be repeated a number of times, to safeguard against the accident of random multipliers much smaller than one paired with those outliers (press F9 repeatedly). There are no analytic formulas for the expected maximum of more than five random variables, so simulation is the appropriate demonstration tool.

Table 6: Alternative severity measure, using PiN proportions from six sectors

Betti and Verma weighting scheme

Aggregate deprivation level: 0.4000

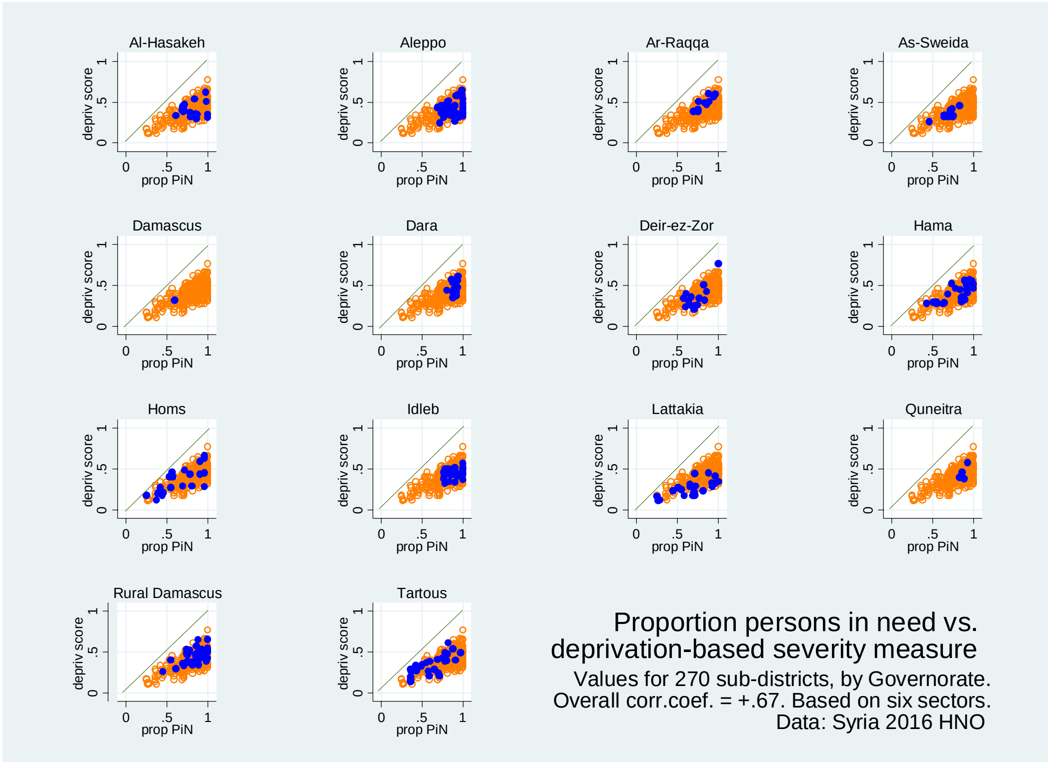
Deprivation weight and contribution to total, by sector

	mean_Prop	Weight	Contrib	Share
propPIN_FS	0.4696	0.3658	0.1718	0.4295
propPIN_NFI	0.2858	0.1491	0.0426	0.1065
propPIN_Shelter	0.1320	0.1609	0.0212	0.0531
propPIN_WASH	0.7125	0.0909	0.0648	0.1620
propPIN_ER	0.5773	0.0839	0.0485	0.1212
propPIN_CCCM	0.3422	0.1493	0.0511	0.1278
Total		1.0000	0.4000	1.0000

The table reports, for each sector, its mean PiN proportion, indicator weight, contribution towards the aggregate (= mean proportion multiplied by the weight), and the share (contributions standardized to sum to 1). The weights are of the greatest interest from the theoretical view point – sectors less strongly correlated with others (i.e., contributing more new information) and/or those with proportions with higher coefficients of variation (i.e, discriminating better) have higher weights. From the policy viewpoint, the shares are important: Food security has the by far highest, due to its relatively high mean PiN proportion *and* its high weight. Shelter is almost irrelevant in this diversity-driven measure.

The new measure has a mean of 0.40 and a range from 0.11 to 0.77. Its distribution is of interest only when we correlate it with the original aggregate PiN measure.

Figure 11: Comparison between the PiN proportion and an alternative severity measure



The alternative severity measure, although derived from PiN proportions, is more conservative than the inter-sectoral PiN estimates. The latter are bunched up towards the maximum (the x = 1 vertical line); the alternative is symmetrically distributed around 0.4. All of the points lie below the line of equality.

Yet, the overall correlation is considerable, if not very strong (Pearson's = +0.67), indicating that they somewhat agree which sub-districts are more severely impacted, and which less so. Readers familiar with the governorates may recognize familiar patterns – the clumping on the inter-sectoral PiN measure in Aleppo, Dara, Idleb, etc. – and may be surprised by the greater dispersion of the alternative measure in the same governorates. Overall, the alternative measure offers better discrimination[15]. But, and this may be reason enough for many users to drop it, it offers no estimate of the overall number of persons in need.

Proxy indicators for persons-in-need

One may be tempted to get rid of unwieldy PiN estimates altogether, replacing them with some category of persons that is commonly understood and is not already mixed up with severity measures by design. In this line, the demography of victims seems fertile. Victims are defined by events (physical or social acts), not by severity measures although they may be part of such measures. Counts of particular victims are conceptually simpler than aggregates of persons in need across sectors, and simpler than transformations of severity

[15] It has higher values than the inter-sectoral PiN measure on all classes of the Generalized Entropy inequality index (Jenkins 2008), e.g. on GE(1) 0.046 vs. 0.028.

ratings. For example, the proportion of IDPs in an area may appear to express the imbalance between needs and resources. Host communities have ideas of varying precision and accuracy of how many displaced persons are in their midst, but some estimates generally exist and can be related to pre-crisis population figures.

However, such simple one-categorical counts or proportions rarely yield valid proxies for persons-in-need estimates. The longer the crisis lasts, the more muddled categories can get – in Syria, for example, with multiple displacements of the same families, the lines between IDPs and returnees have been blurred. Involuntary movers comprise not only IDPs, but also refugees and some returnees. And it is not only the involuntary movers that have acute needs; so do the involuntary stayers, particularly in besieged areas.

One could, of course, try to be creative and combine some more refined categories (as we did with the distinction between moderate and acute needs). One large survey in Syria went as far as eliciting from village and town-level key informants estimates of households still living in the same pre-war dwellings[16].

The proportion of such households in the local community does have proxy indicator potential. We could claim some face validity if we attempted to express severity as the multiplicative effect of displacement and needs/resource imbalance. A possible measurement model might look like this. The first term measures the proportion of the population displaced at least once. The second captures the ratio of all people (incl. IDPs) in acute need to those left with more resources. These, supposedly, are the ones still in their old – howsoever badly damaged - residences:

Proxy for proportion of persons in need =

*[1 – (Households still living in their pre-war dwellings) / (Pre-war households)] * [(Current total households) / (Households still living in their pre-war dwellings)]*

The range of this expression exceeds 1. It is not defined if no one is living in their pre-war dwellings.

It would be easy to assail the validity of the measure on various grounds. These range from changes in household composition to effects of multiple displacement as well as variations in structural damage and service disruption, none of which it takes into account. But this should not discourage experiments if the data already exist. Exploring them might give insights into the finer contours of the severity map, below the sub-district – insights potentially useful for the fine-tuning of relief and protection.

Larger doubts arise over the reliability of such estimates. The number of households still occupying their pre-war dwellings may be harder to estimate in larger communities, and in small communities relatively minor absolute errors in the denominators would cause large errors in the expression.

[16] "Dwelling" is more flexible than "residence" and includes households that changed floors in a partially destroyed building, or move to another building in a multi-building family compound.

Persons-in-need estimates – on balance

If one goes that far, then why not take the next step – moving away from the persons-in-need based approach altogether? Surveys that are able to extract something as specific as information about households still living in their pre-war dwellings should typically also be the ones that collect a good number of other indicators anyway, in various fields. These can be bound together in process and measurement models that produce severity estimates without dependence on persons-in-need estimates.

But it is unlikely that the humanitarian community could ever dispense with persons-in-need estimates. They serve a purpose, both internally and externally. Internally, response plans are difficult to imagine in the absence of guidance that determines not only the qualitative – what kinds of persons should be paired with what kinds of resources and activities -, but also the where, when and how many. In other words, these estimates are needed for priority-setting.

Externally, persons-in-need figures are a key element in communicating with the political systems that fund the humanitarian effort. The presumption is that there are people out there deserving of assistance and protection; they can be contacted, counted and helped – these actions, if done effectively, have a certain unit cost – and for this purpose those in need receive a temporary client status in the system of humanitarian organizations (Hasenfeld 1972). Their estimated number is the hinge between the cost of reducing the suffering and funding requests (and subsequently the received budgets). Without this link, only market-based solutions would remain – arrangements that determine price and quantity autonomously. Many of these are morally or politically not acceptable.

The way forward is to create coherence, transparency and discipline in the relationship between persons-in-need estimates and levels of severity. Not everyone is equally severely affected. Measures of unmet need must discriminate. If (almost) everyone is in at least moderate need – a reasonable assumption for long-drawn conflicts and some complex emergencies -, then "acute need" no longer discriminates enough. Additional discrimination can be created in two ways:

- Either sharper distinctions are created within "persons in acute need" – "persons in critical need" and "persons in catastrophic need" is one of many ways of naming them. These must be applied with restraint lest they wear off in repeated grade inflations.
- Or, if that is morally, politically or technically not feasible, the estimates of those initially classified as "acute" are split on the strength of some severity index based on other indicators. Areas with higher index values would see a larger proportion recognized as acute; the rest would be included with those in moderate need[17].

Finding a defensible function for such a re-allocation would be difficult; any one would likely be contentious. It would not meet the needs to the individual sector coordinators, but could ultimately result in better combined estimates of persons in need.

[17] Technically, the initially estimated proportion of persons in acute need can be regressed on the index, possibly with a non-linear term. The predictions would be used as the revised proportions, with the proportions of those in moderate need adjusted.

A radical alternative

Our extended discussion of severity ratings and persons-in-need estimates has made one thing clear: The two wings of sector-based model do not cohere.

- As practised by the Syria 2016 HNO, one severity rating is given out per sub-district, as a synopsis of multiple sub-criteria of highly variable specificity. This cognitive complexity could not be dealt with directly. The assessors instead relied on a process of anchoring and adjustment.
- The persons in need were estimated through rather vague expert judgment processes that took into account the severity rating, but also specific information that the sector coordinators collected and linked in their own professional perspectives. Some sectors restricted the variability of the PiN proportions across the sub-districts.

The WoSA and HNO in Syria were massive exercises in data collection and analysis. It is unlikely that similar efforts elsewhere would achieve better coherence. It may be time to consider a more radical approach. This would import the severity levels into the breakdown of the population in need itself. This means that the sectors would cooperate in estimating, for each of them, the number of persons in an area in catastrophic need, critical need, etc.

This would generate more conservative estimates, with a requirement to document reasons how the chosen population proportions were determined. Although six or seven levels would be too many, a simplified elicitation format could be thought out. The local partners would be asked to distribute the latest population estimates over the severity levels; these would be uniformly defined by the associated survival chances for sectors that cater to material needs, and by yet-to-be-determined criteria for the other sectors. As with other expert judgments, a most plausible estimate, a minimum and a maximum could be requested. This table offers a fictitious scenario for an area.

Table 7: Severity estimation through graded persons-in-need estimates

Levels of need in the _______ sector	Population and persons in need	What is the evidence for your estimate?
The estimated **total** population currently in ___________ Area	100,000 (90,000 – 120,000)	
Those with **minor to major** problems regarding the needs in this sector [will survive without assistance, but longer term damage likely]	60,000 (50,000 – 70,000)	
Those with **severe** problems [will survive if assistance given within one month]	30,000 (25,000 – 35,000)	
Those with **critical** problems [will survive if assistance given within a week]	7,000 (5,000 – 8,500)	
Those with **catastrophic** problems [most will die even if assistance is available today]	3,000 (2,000 – 4,000)	

Note: This table is identical with Table 1 in the summary.

The nutshell evidence notes would allow coordinators to review the numeric claims for minimal plausibility. Aggregation would be additive in two dimensions – areas and levels. Combined sector PiN estimates for a given area would take the maximum values, starting at the catastrophic level, and working its way down, until the cumulative total fills up the total population. Thus, although all sectors may initially report positive numbers in the lower categories, they may be zero for the combined PiN estimates, due to the maximum estimator. Once the combined estimates have been established for each area, geographic aggregation is purely additive.

The uncertainty of the estimates could be computed by simulation, but at this purely notional point, that is of minor importance. The key point is to offer a severity measure that is ratio-level from the start, makes hard choices unavoidable, and demands documented reasoning. On the strength of the evidence, sector coordinators could override numeric claims from partners in the field, and the coordinating body would critically review the sector submissions side by side.

If carried out in mutual discipline, such a scheme might bring severity, priority and victim demography closer together.

The indicator-based model

The indicator-based model enjoys greater flexibility than the sector-based model does. It is not beholden to sectoral interests, perspectives and data holdings. It can arrange available indicators to underlying concepts in exploratory ways and can choose among several methods for aggregating indicators into sub-indices and indices. It has greater freedom to incorporate pre-crisis indicators.

This freedom comes at a price. The burden to prove that measures validly express severity is heavier. Indicator standardization, weighting and aggregation all need justification because each step offers alternatives. If the analyst is far removed from the agency that defined and measured a particular indicator, she may misunderstand its meaning, limitations and tacit local adjustments, possibly to diminish the validity of the absorbing measure. And, as already noted in the previous chapter, even the most valid and reliable severity index cannot directly be translated into persons-in-need statistics.

As noted in the introduction, we justify the severity measure by building a process and a measurement model and then reviewing assumptions and results at each step. Typically, we will be led to build the process model around concepts of intensity, exposure and pre-existing conditions (vulnerability). Each of these needs a construct, a composite measure that incorporates the important conceptual aspects. The measurement models define how we get from the indicators to the composite measures – or, if the indicators do not yet exist, how they might be collected or replaced by expert judgments or simulated values.

Process models

Process models, also called "structural models", define relationship between concepts. Beyond mere conceptual models, in which boxes, each with the name of a concept, are linked by arrows, process models move closer to fully formulated quantitative models. They assume that for each concept a construct exists – a variable to which measurement operations subsequently assign values. The process model links the constructs by operations that are elementary arithmetic operations – addition, multiplication, etc. – or take functions (e.g. a logarithmic transformation) of one or all constructs and links the transforms. The process model, in other words, wires concepts together through defined quantitative input-output relationships, then waits until the measurement models and actual data supply the inputs, in order to determine the outputs.

The notional model with six terms advanced for the Global Severity Index (see page 19):

$$\textit{Severity} = \frac{\textit{Vulnerability} * \textit{humanitarian outcomes} * \textit{duration} * \textit{threats}}{\textit{Humanitarian access} * \textit{capacity to cope and respond}}$$

is somewhere in the grey area between a conceptual and a process model ("no humanitarian access" cannot be zero; a transformation is needed). The reduced model

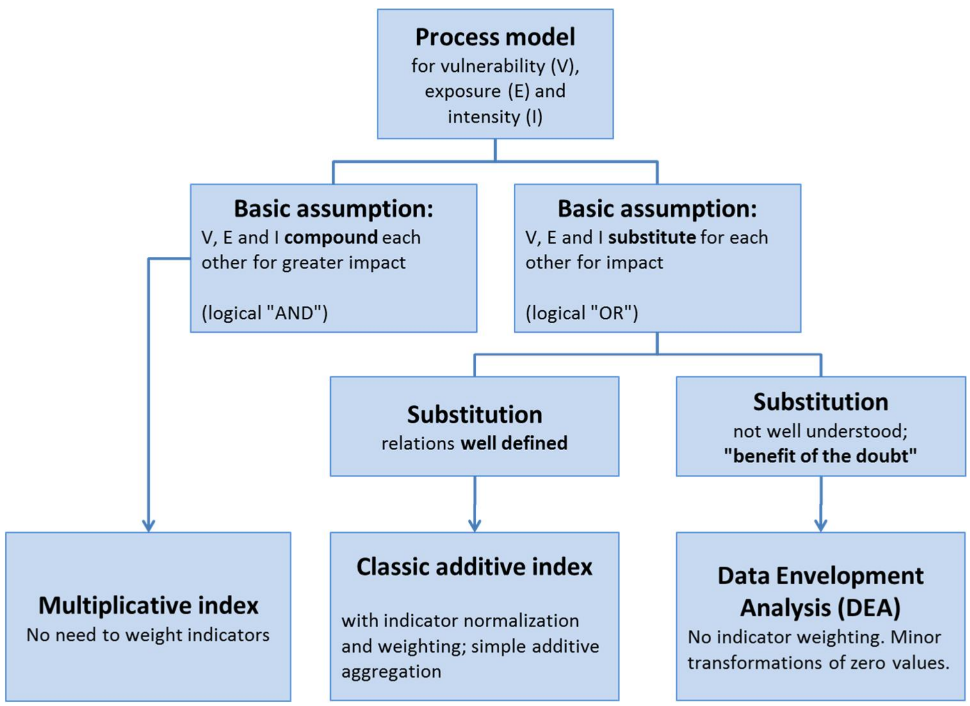
$$\textit{Severity} = \textit{Intensity} * \textit{Exposure} * \textit{Vulnerability}$$

is a process model. The numeric values that three measurements supply will be plugged into the equation; the product from the multiplications is the severity measure. The measures should be dimensionless; the measurement models should take care of that (which may be a hard-to-overcome challenge – think of the Mercalli Index measuring earthquake intensity – what transformation would be adequate so that for given exposure and vulnerability a reasonable severity level would be proportionate to the intensity?).

A major distinction

A major distinction can be made between models whose constructs compound each other in their effects on severity, and those that substitute for each other. The first lead to multiplicative aggregation, the second come in different flavors. Additive indices and a procedure known as Data Envelopment Analysis belong here.

Figure 12: Major types of process models



The three types are not exhaustive. Researchers have proposed other aggregation types (for an example, see Munda and Nardo 2009). However, the three that we discuss here can be implemented with relative ease in Excel, respectively in a user-friendly DEA freeware that exports to Excel (Benini 2015b) – practical considerations that win over most humanitarian information managers.

Measurement model

General considerations

Rarely discussed in the humanitarian data management world, the ways indicators are related to concepts or constructs can differ with important consequences. In severity measurement, we may not notice this because we believe the work is done when we are done combining a number of indicators in some composite measure that expresses severity.

A key point concerns the relationship between *indicators whose values we observe* and the underlying *concepts that are not directly observable*. Our research constructs give these concepts their quantitative expressions. An important distinction has to do with the direction of causality. Is the construct the cause, and are the indicators effects chiefly of this one cause? Or are the indicators the causes that converge onto one common effect captured as a construct?

In practical life, assigning constructs and indicators to either cause or effect is not always possible. In empirical research, however, there are two questions the answers to which take us further:

- Are the indicators expected to be correlated, or not?
- Does adding an indicator to, or dropping it from, the measurement model alter the meaning of the construct, or not?

We discuss this with two fictitious vignettes:

1. **From indicators to construct**: Inhabitants of a coastal area are vulnerable to hurricanes and attendant surge floods. Rich people live in flood-resistant concrete buildings. Poor people live in less sturdily built wooden structures and are more vulnerable. We measure vulnerability as a composite of the local poverty rate and elevation above sea level. There is no ex-ante reason to expect that higher poverty and lower elevation go hand in hand. However, excluding, in this situation, either poverty or elevation would invalidate the vulnerability construct. Ergo, the causality flows from the indicators to the construct.

2. **From construct to indicators:** A civil war wrecks destruction and deprivation on local communities although to greatly variable degree. We measure the deprivation through levels of unmet needs in several sectors. The needs too are not directly observable; we infer them from indicators such as the price of the staple food (food security), the price of drinking water (WASH), the proportion of defunct health care facilities (health), and others. It is reasonable to expect that levels of unmet needs are positively correlated across sectors although the strength of correlations can vary. The inability to find good indicators for some particular sector does not change the meaning of deprivation. The causal flow is from construct to indicators.

Researcher have named the first measurement model “formative”, the second “reflective” (Fornell and Bookstein 1982, Jarvis, MacKenzie et al. 2003). In the first, all indicators are necessary to “form” the construct. In the second, a common factor is “reflected” in multiple indicators. In the context of severity measurement, these are important distinctions and consequences:

Table 8: Formative and reflective models in measurement

Aspects of the model	Formative	Reflective
Direction of causality	From indicators to construct	From construct to indicators
Examples	Poverty and geo-physical location define vulnerability	Societal capacity loss (e.g., as a result of war) causes unmet needs to surge
Correlations among indicators	We cannot expect that the indicators are correlated. Absence of correlation does not imply validity or reliability problems.	We expect that they are correlated. Absence of correlation signals a likely problem about the construct or the measurements.
Dropping an indicator from the measurement model	May alter the meaning of the construct, may invalidate it.	Does not alter the meaning of the construct.
Is completeness required?	Requires the study of all essential aspects of the construct, and for each aspect an indicator should be found.	May have selective indicators

Note: Largely following Jarvis et al., op.cit.: 201

The practical consequences that we must observe concern: 1. When the absence of correlations between indicators should alert us to potential validity (indicators do not express concept) or reliability problems (indicators have measurement error), and conversely when the absence is innocuous; 2. The need to represent all essential aspects of the concept in the indicators, vs. greater freedom to be selective.

Specific measurement issues

Indicator-based severity models are faced with a number of measurement issues most of which are familiar from composite measures in general. We can broadly subdivide them into issues that arise from uncertain and incomplete measurement and those related to the metrics of indicators.

Forms of uncertainty

Uncertainty and incompleteness present themselves in multiple flavors; their remedies are more or less effective, depending on wider constellations:

- Measurement error, in theory, is mitigated by the inclusion of several indicators in each construct. Practically, good indicators are in short supply. Adding, merely for the sake of diversity, any indicators known to have large errors does not help.
- Missing data may be filled with various imputation techniques, but at a much more basic level one simply wishes that codebooks, data entry personnel and analysts distinguish between missing values and genuine zeros.
- Severity indices may seem more robust than they really are in the face of measurement error and of arbitrary weights and aggregation modes. Robustness can, and often should, be simulated by varying error components, weights and aggregation modes; at least for the latter two considerations, limited checks are fairly easy to perform in spreadsheet programs.
- Indicators sometimes arrive in different granularity. This problem, known as "multi-resolution", may affect a single indicator or the relationships between indicators:

- For example, for district X, estimates of damaged buildings may be available for every sub-district whereas in Y only a district-wide estimate has been advanced.
- For the second situation, suppose that we want to combine two indicators of pre-existing conditions, say, the poverty rate and the proportion of persons with disabilities. We may find that surveys were conducted on both groups of people, but estimates subsequently were published for different administrative levels[18].

A detailed treatment of uncertainty and incompleteness would exceed the space of this note. Moreover, by the time the analyst begins the construction of severity indices from available indicator datasets, many of the choices affecting uncertainty have already been made at the earlier stages, during assessment design, data collection and acquisition of secondary information.

The **metrics** are commonly addressed under the title of “measurement levels”. Many readers may be familiar with the time-honored distinction among nominal, ordinal, interval and ratio-level measures, and with the kinds of transformations that are admissible at each level. This classification has been criticized (Velleman and Wilkinson 1993). An alternative with six levels (Mosteller and Tukey 1977):

1. Names
2. Grades (e.g. freshmen, sophomores etc.)
3. Counted fractions bound by 0 and 1 [also when expressed as percentages]
4. Counts (non-negative integers)
5. Amounts (non-negative real numbers)
6. Balances (any real number)

is considered more valid, if less familiar. For example, in the traditional classification, percentages and proportions are considered ratio-scale and thus, in theory, admit the same kinds of scale transformations as, say, between Fahrenheit and Celsius. It is obvious, however, that proportions cannot be arbitrarily scale-shifted. It is worth discussing some critical precautions for each of those six levels.

Names, or nominal indicators

Names imply a set of dichotomous variables. Each variable designates the elements that share a given name and those that do not (their complement). Thus, the villages of a crisis zone that consists of Areas A, B and C, have three properties related to the names of the area; a village “is part of A” or not, “part of B” or not, “part of C” or not. Trivial as this may seem, one can be grateful for statistical programs that automatically generate from a multinomial variable a set of dummy (0, 1) variables, convenient for further specific analyses. In Excel this is a bit more involved, requiring IF-functions.

“Is part of Area A”, of course, is unlikely to be useful in a severity index. We expect nominal indicators to be substantively meaningful. For example, if the process model calls for a

[18] Benini, Chataigner et al. (2016) discuss the multi-resolution issue in the context of gaps in needs assessments, see http://www.acaps.org/sites/acaps/files/resources/files/info_gaps.pdf.

hazard sub-index, the fact that a government or NGO program designed to combat the hazard is active in the area may be a good proxy when other, more direct measures are not available. Technically, this too is expressed in a binary indicator, taking the value “1” if the program is active in an area, and “0” otherwise.

Binary indicators are information-poor – they only take two states. One may want to combine several of them in a sub-index. This can happen in different ways:

- If we have good reason to consider some indicators more important than others, we can give those higher weights, then aggregate the weighted ones additively. Note, however, that the weights also multiply possible measurement error.
- If the indicators all have the same direction (e.g., the value “1” in every one points towards greater severity) and are of equal importance, we may simply sum them. Summing implies that all have the same weight.
- If we believe that indicators with lower frequencies express increasingly severe conditions (because in a particular context rarer conditions, where they are found, point to more severe underlying problems), we may increase weights in step with decreasing frequencies. For sufficiently many such indicators (e.g., of services that were functional before the crisis, and are now absent), we may let a statistical algorithm (e.g., item response models) search for a common scale (e.g., of institutional collapse).

Grades, or ordinal variables

In severity measurement, ordinal variables have traditionally arisen from one of two operations:

- the ranking of originally interval- or ratio-level variables (or, in the alternative classification above, of counted fractions, counts, amounts or balances)
- the elicitation of expert judgement and key informant opinions that resulted in ratings and ratings.

The first is motivated by the convenient manufacture of a presumed common metric across indicators. It causes uncalled-for information loss and produces, in the aggregation of several rank variables, invalid, non-sensical results. This behavior has been rampant in humanitarian information management for some years; it appears to be abating. Benini and Chataigner (2014, op.cit.; see resources above) discuss this extensively. The short version is: Don’t rank.

In the second, rating and ranking mean mapping a set of elements (verbally described) to a set of consecutive numbers, usually starting in 1:

- Ratings express an *order on a scale*; an element is assigned to the scale level whose meaning it fits best. It is rated without regard for the other elements. The severity ratings discussed in the previous main section (pages 22 sqq.) exemplify this operation.
- Rankings are *comparative*; an element is ranked higher than another if it exceeds the other in the relationship that defines the ranking.

Ordinal variables that express **ratings** are generally more challenging to aggregate. One can take rank-based statistics such as the median or the maximum. Suppose key informants, each speaking about a sub-district, rate ten services on a five-point scale of functioning. The row median measures the overall service level in the sub-district; sub-districts can thus be compared. The column median expresses the relative functioning of a given service compared in all areas; services can thus be compared. These medians, however, may not be very sensitive to extremes, an aspect that the assessment needs to describe. One will therefore endeavor to transform the scale to a higher measurement level. Different transformation may be feasible:

- counting the number of indicators on which the sub-district in point was rated at some levels (e.g., the lowest or second-lowest levels),
- the *ridit* measure demonstrated earlier (pages 25 sqq.),
- a Likert scale (Wikipedia 2011b) if the assumption of equal distances between levels is plausible (not likely!), or
- some kind of statistical item response model (Wikipedia 2016c), for which the scale may have to be recoded to fewer levels in order to obtain robust estimates.

Another alternative consists of assigning each rating level a ratio-level weight motivated by the policy context. For example, the severity index that UNOCHA created for low-level communities (Village Development Committees, VDCs) affected by earthquakes in Nepal in spring 2015 incorporated an access indicator. It was based on the means of transport that were feasible at that time:

Table 9: An ordinal scale of access to village in Nepal, converted to ratio-level

Means of transport	VDCs	Weight
No access by any means	33	0.00
Access only by helicopter	85	0.20
Tractor access	107	0.50
4x4/Pickup access	124	0.80
Truck access	279	1.00
N / mean	628	0.71

The rating scale itself here (from 1 to 6) is no longer important; ratio-level weights have been assigned; their arithmetic mean is a legitimate operation. Regions can thus be compared by the scale means for their VDCs. The scale itself could be understood as an aggregation of binary indicators. We assume that if a VDC can be accessed by a heavier road vehicle, it can be accessed also by all the lighter types as well as by helicopter. The weights on the binary variables would be 0.20 for helicopters and trucks each, and 0.30 for tractors and 4x4 pick-ups. Additive aggregation produces the listed weights of the ordinal levels, as in the table.

The transformation of **rankings** to interval- or ratio-level variables has a long tradition. It proceeds in analogy to an election system. Experts or key informants “vote” for items in the unit that they represent. They state which of M items (generally problems, needs, humanitarian response options) is the most important, which the second most important, etc. They rank $K \leq M$ items, and this happens in all N units (villages, districts, social groups, etc.). The most important item receives a score of K, the second most important $K - 1$, etc. If $K < M$, the non-ranked items receive scores of 0. The sum of the scores for an item, over all units, is known as its Borda count (Borda 1781, Wikipedia 2011a). As count variables, means of Borda counts are legitimate; sets of units (e.g. the villages of two regions) may be compared on them.

Be aware of the direction in which the Borda count aggregates. The experts or key informants rank items in their own units only. In election terms, the items are the candidates, the units are the voters. As a result, the Borda count is a measure pertaining to an item, not a unit – items can be compared to other items for their relative importance or severity, and subsets of units (such as districts) can be compared on the Borda counts of a given item. But there is no aggregate measure for an individual unit over all items – village A is not more or less important, affected etc. than village B – the voters are all equal.

The interpretation of the Borda count – or, more intuitively, of the Borda count divided by the number of voting units, the mean score – has to be prudent. If a voter does not rank all M items, but only the K that he/she prefers most, $M - K$ items receive a zero score. However, for the true importance or preference that the Borda count estimates, the difference between adjacent ranked items (e.g. the first and second most important ones) and that between the last ranked one and any unranked ones may not be the same. In a $K < M$ Borda count, therefore, the average item scores should be interpreted conservatively, as interval-level values. This means that items can be compared by the differences of their average scores, but not on ratios of scores.

To illustrate, assume that voters each ranked the three most important interventions in a list of ten. It turns out that vaccinations attained a mean Borda score of 1.6, road repairs one of 0.8, farming tool distributions one of 0.4, with the other seven sharing an insignificant 0.2 votes. We *cannot* claim that vaccinations are “twice” as important as road repairs. We *may* say that the difference in importance between vaccinations and road repairs is far greater than that between road repairs and tools, and hence all the others. If, however, $K = M$, a ratio-scale interpretation seems defensible, and so then are ratio comparisons[19].

[Sidebar:] Borda count analysis of coping strategies in Bangladesh floods

This example is for a Borda count analysis for $M = 15$ items, of which every voter ranked $K = 5$. The voters in this rapid assessment (ACAPS and et.al.; 2011) were 59 focus groups drawn from flood-affected communities in Bangladesh; the items are 14 distinct coping strategies plus “Other strategies”. The table sorts the strategies by decreasing mean Borda score for all focus groups; the mean scores are broken down by type of temporary living arrangements of the group members.

---

[19] Data management of ranking data in the Borda count perspective for Excel users is discussed in , and a demonstration workbook is available at,
http://www.acaps.org/sites/acaps/files/resources/files/heat_maps_as_tools_to_summarise_priorities_2011.pdf
http://www.acaps.org/sites/acaps/files/resources/files/heat_maps_as_tools_to_summarise_priorities.xlsx .
For a recent example of the Borda count employed in rapid assessments, see Limbu et al. (2015).

Table 10: Borda count example

Coping strategies	Living arrangement: Collective centers	Roadside / embankments	Marooned	Home undamaged	Damaged or water-logged	All
Reduce meal size	2.88	4.13	2.43	3.13	1.92	2.88
Borrow money at high interest	2.75	1.44	1.57	2.88	2.77	2.18
Purchase food on credit	2.00	2.00	2.29	1.38	2.31	2.05
Sell livestock and poultry	1.75	1.63	1.57	2.25	2.31	1.90
Eat less preferred food	1.00	2.13	1.43	2.63	0.62	1.52
Out-migration of household members	0.25	0.56	0.93	0.88	1.31	0.80
Sell household utensils/utilities	0.88	0.06	0.71	0.75	0.46	0.50
Sell land or trees	1.00	0.56	0.21	0.25	0.31	0.48
Female adults restrict food consumption to feed the children	0.50	0.13	1.07	0.00	0.54	0.47
Sell labor in advance	0.63	0.44	0.50	0.63	0.23	0.47
Eat wild foods like roadside vegetables etc.	0.75	0.44	0.21	0.88	0.31	0.45
Male adults restrict food consumption to feed the children	0.00	0.19	0.79	0.00	0.46	0.33
Send children to friend / relative's house	0.13	0.81	0.29	0.00	0.00	0.30
Send children to work	0.38	0.31	0.29	0.00	0.38	0.28
Other	0.00	0.00	0.00	0.13	0.23	0.07
Groups interviewed	*8*	*16*	*14*	*8*	*13*	*59*
Concentration of priorities (Standard deviation of item Borda counts)	*0.94*	*1.12*	*0.76*	*1.13*	*0.92*	*0.88*

Legend:

Priority

- Very high
- High
- Intermediate
- Low

Source: See footnote on page 53.

The main message is that 5 out of the 14 coping strategies clearly dominate (the drop in mean scores between items # 5 and 6 is steep); four of them directly concern food. Because of the small sample, comparisons among the five groups have limited value. That roadside and embankment dwellers were facing greater difficulties in accessing credit is plausible. Frequently it is the very poorest who get stranded in this type of location.

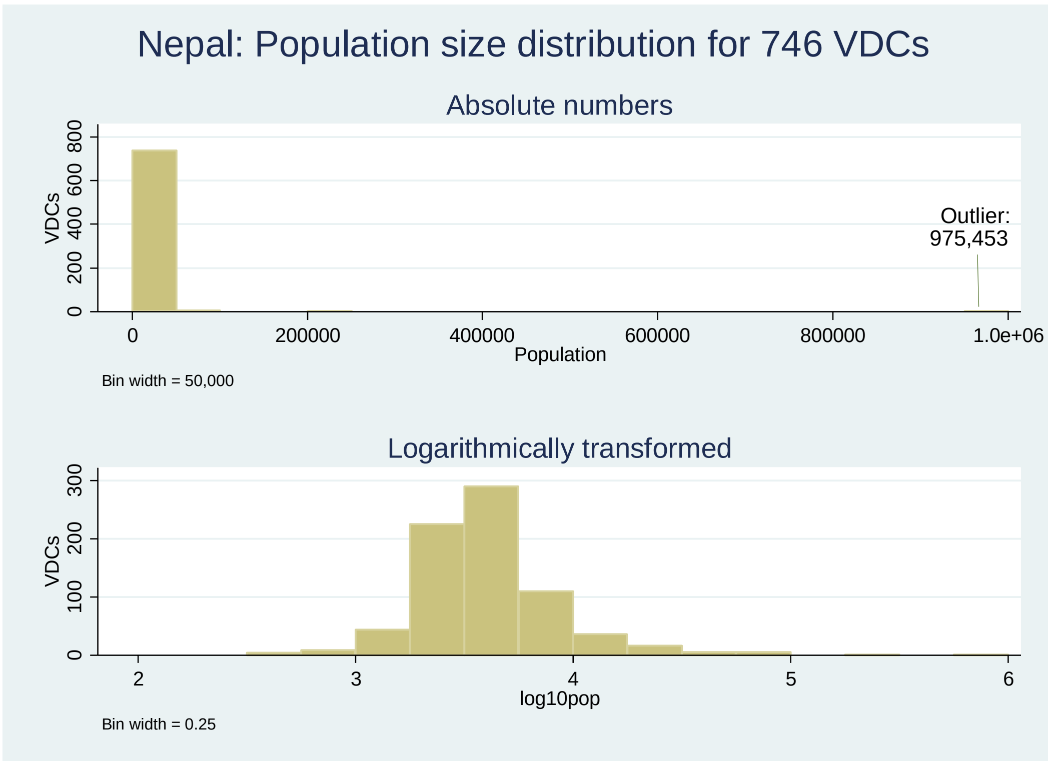
Counts

Counts are variables with integer values; they are ratio-level, with the exceptions that they cannot take negative values.

Population size, either as individuals or as households, is by far the most common such variable in the severity context.

Count variables, including population, often have very skewed distributions. For graphing and for certain statistical purposes, it may be helpful to work with their logarithmic transformations. This graph illustrates the issue with the distribution of the populations of Village Development Committees in Nepal.

Figure 13: Population size distribution - absolute vs. logarithmically transformed



In the absolute-value representation, a very large outlier causes most other population values to cluster in the first bin. This histogram carries little information; we only learn that few VDCs had populations larger than 50,000. The histogram of the log-transformed population lets us immediately see that most VDCs had populations larger than 10^3 = 1,000 and smaller than 10^4 = 10,000. The outlier, of course, remains an outlier in the logarithmic transformation as well, but its effect on the overall distribution is milder.

[Sidebar:] Challenges with event count indicators

Special vigilance is needed with *event counts*. Such count variables record incidences – the number of events of a defined type counted for a given unit in a given period of time. Often these occurrences would be seen to fluctuate wildly from period to period – if we observed them during more than one period. The wilder the fluctuations, the less suitable are these variables as indicators of stable properties of the units or objects that they are supposed to characterize.

In such situations, two questions should be asked:

1. What processes cause the distribution? As statisticians would say: What is the "data generating process"? Note that, yes, we may consider this variable to be part of a measurement model (an indicator within a sub-index), but we need, notionally, a kind of process model for itself. What causes high and low values in this variable?
2. Are there particularly many, or particularly few, units with zero events recorded? Are there special mechanisms responsible for the frequency of zeros?

The zero observations are consequential. An excess of zeros may result either from error, or from the presence of multiple causal mechanisms:

Error: Was the unit closely observed, and yet not a single event was noted? (= no error). Or was it somewhat observed, but so superficially that events that did occur were never noted? (= measurement error). Or was it not observed, and a missing value should have been entered, but, instead, a zero was recorded? (= processing error).

Example: The above-mentioned severity assessment in Nepal collected statistics on the number of landslides that had devastated VDCs since the earthquakes, many of them finally triggered by monsoon rains. Out of 746 monitored VDCs in 14 districts, the mean was 10.7. All districts but one reported some slides. In over half (416) of the VDCs, the recorded number was zero. Among the 330 VDCs with reported landslides, the number ranged from just 1 to an astronomical 555, with a mean of 24.2.

Side by side with this statistic, the number of VDCs with zero landslides seems excessively high[20]. We must assume that it was not possible to monitor landslides reliably.

Multiple causal mechanisms: Are the events driven by a mechanism outside of, or supplementary to, the drivers of severity that we included in the process model? Or, conversely, is there a special mechanism at work that suppresses the events that otherwise we expect at significant rates?

Example: In assessing the severity of Typhoon Haiyan impacts on 408 municipalities in the central Philippines in 2014, the Protection Cluster experimented with some novel indicators (Benini and Chataigner 2014:53 sqq.). One of them was the count of incidents since 2010 that were attributed to an armed opposition movement. Reports of such incidents had come from 77 municipalities; these 77 were scattered in 10 out of the 16 provinces in which the assessment was conducted. Three provinces had significantly higher incident counts than the others.

Philosophically, the inclusion of this indicator in the severity index is problematic; the activism of armed groups is a *causal* factor whereas the index otherwise was formed on indicators of *consequences* (of the typhoon and indirectly, in the poverty rate, also of civil unrest). Mixing indicators from these two types sits awkwardly with the desired coherence of the process model.

Here, however, we are interested in the distribution of zero counts. It is straightforward to posit a separate mechanism from the typhoon, notably because part, if not most, of the incidents took place before the storm. In provinces where no incidents were recorded, one may assume that the armed opposition was not active. In provinces with incidents, all municipalities may have been at risk of this kind of violence, but the armed opposition chose not to strike in some or most during the counting period. The distinguishing factor then is the presence of the armed opposition, measured at the province level. Zero counts for municipalities in provinces without incidents result from one

---

[20] A statistical model (zero-inflated negative binomial regression, with population as exposure and the districts as inflation factors) predicted significant zero inflation (Vuong's z = 8.29). The probability for a VDC to always report zero landslides (because it is in a district with few or none) is 0.40, the observed proportion of VDCs with zero slides is 0.56, the probability to report zero including also random variability and population (exposure) is 0.59.

causal circumstance (absence of the opposition); zero counts in provinces with incidents are random realizations of a different process (active opposition).

Practically, what is one to do with such indicators?

- If the observed number of events has a very skewed distribution (as in the Haiyan case), it should be transformed. Because of the strong fluctuations (which we would likely see if we had data for more than one period), the variable should undergo a so-called variance-stabilizing transformation. Counts often follow a Poisson distribution; if so, the Freeman-Tukey transformation (Wikipedia 2014):

$$x \rightarrow x^{0.5} + (x + 1)^{0.5}$$

  may be appropriate. Thus, if the count of events is 50, it transforms to sqrt(10) + sqrt(11) = 6.48. Depending on the analytic context, the transformed indicator may still have to be standardized, such as by dividing by the population in the unit.

- In addition, if one can pinpoint a subset of units in which the special mechanism suppresses the events (e.g., municipalities in provinces from which the armed opposition is absent), a special binary indicator may be created, taking the value "1" for units at risk, and "0" for those not at risk.

The inclusion of these two indicators in one or the other of the constructs that make up the main process model has to be justified. Are they more intimately related to the exposure, the intensity, or to the pre-existing conditions? The weights, too, are similarly tricky to decide. We could let the Betti-Verma algorithm (explained in detail by Benini and Chataigner (2014)) figure them out, together with the weight of the other indicators included in the same construct.

Those are technicalities with which the reader may deal once he is facing event incidences as candidates for severity-related indicators. More fundamentally, he should first ask:

- What causes some units to experience few events, and what others to have many?
- What causes some to have zero events, and others positive numbers?
- Are there different mechanisms at work, and can they be identified?

If event count indicators are key to the construction of a severity index, one should consult with epidemiologists. The analysis of incidence data is part of their core expertise.

Amounts and balances

These kinds of variables – amounts with values of zero and larger, balances with positive and negative values – are, in the traditional classification, either at the interval or even ratio levels. Further above, we have argued that certain priority measures – the mean score in the modified Borda count – should be regarded as interval-level; they have no intuitive zero points. Outside such derived measures, it is hard to fancy any indicators with an observable physical basis, continuous metric and humanitarian interest that would not permit ratios.

Ratio-level indicators are ideal for index construction. For severity indices, datasets that offer relevant indicators exclusively at the ratio level may be the exception rather than the rule. This is not always evident, except on close inspection. Information managers often treat ordinal indicators as though they were ratio-level; for example, the food security classification (IPC Global Partners 2012) may masquerade as a ratio-level indicator inside an additively aggregated index (as in Liew 2015).

Let us optimistically assume that, somehow, by legitimate hook and crook, we manage to transform all candidate indicators to ratio levels, or at least to low-level sub-indices with ratio-level character. For the generic options to consider in standardizing, weighting and aggregating indicators, we refer to the ACAPS resources listed above. Here we discuss aspects of the selection of indicators, assuming that we have the luxury of an information-rich environment so that we can possibly exclude some candidates.

Saisana, in a presentation to humanitarian researchers (Saisana 2015), emphasizes the correlations among candidate indicators as selection criteria. A candidate should not be included if its correlations with others are

- Negative (indicating conflicting issues) [i.e., negative after all the candidates have been orientated in the same direction, towards greater severity. AB]
- Random (indicating that de-facto it is independent of the construct to be measured)
- Strong (indicating that aspects of the construct are double-counted).

These prescriptions make good sense generically. They should be followed in constructing low-level sub-indices. Yet, already for mid-level constructs, there may be good sense in admitting uncorrelated or negatively correlated measures, depending on the underlying theory for the guiding concept. Liew (op.cit.), for example, broke his socio-economic vulnerability measure down into four "categories". Two of them, standard of living and labor capacity, are measured by one indicator each – the poverty rate (headcount) and the proportion of the VDC population that migrated abroad. The two are uncorrelated. This is expected because 1. Poorer communities may feel greater migration pressures, but have less developed long-distance social networks; 2. Higher migration later may produce larger remittances, which reduce poverty. The lack of a manifest correlation masks an equilibrium process, locally very diverse, yet overall influencing the pattern of vulnerability.

At the top level of constructs that are joined in the severity index, negative correlations are expected, and the art of severity index construction is to find weights and aggregation functions that produce a plausible severity index distribution. Exhibit A for this situation is the negative correlation between exposure and intensity. The correlation between populations and the severity index values – actually an intensity measure – that Liew calculated for 627 VDCs is a highly significant -0.31 (Spearman's rank order correlation). Not surprisingly, the UNOCHA teams in Kathmandu struggled with the dilemma between the unweighted severity index (= intensity) and the population-weighted flavor (= severity). The former resulted in maps that showed the sparsely-populated high-altitude VDCs to be the far more severely affected communities. Maps of the latter showed the severely affected VDCs more evenly distributed, with some medium- and high-severity clusters in lower-altitude areas.

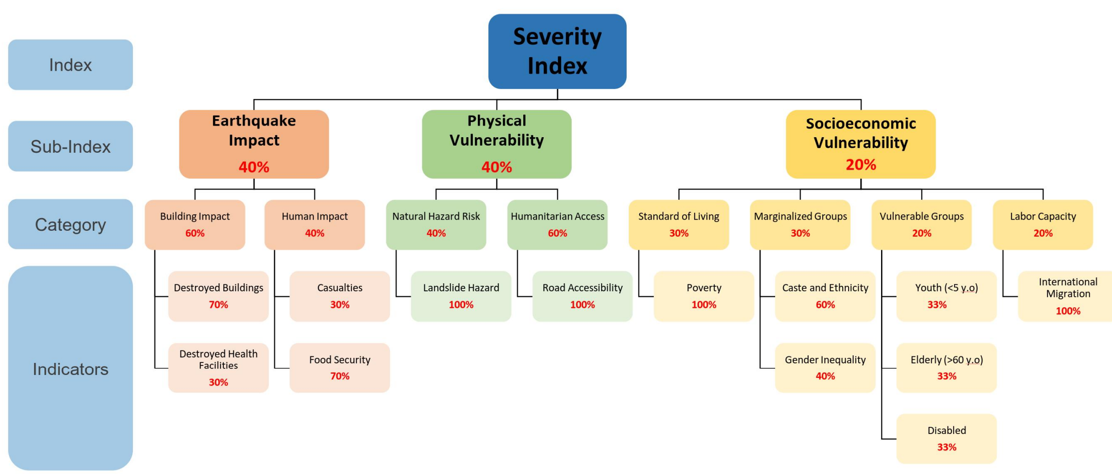
But such is humanitarian life, or at least humanitarian sense-making. If the process model defining severity demands a certain key construct, it should not be excluded on grounds of correlation with other constructs. The analytic challenge then becomes how to aggregate them. For the purpose of this discussion, we want to assume that the measurement of the key concepts - exposure, intensity and vulnerability - has been resolved and is unproblematic. We vary the aggregation mode in response to key assumptions of the process model. This is the topic of the next session, for which we go back again to the Nepal data arranged by Liew.

Alternative severity measures: The Nepal earthquakes of 2015

From the two earthquakes that struck Nepal in spring 2015 emerged a fairly information-rich coordination environment. The European Union's Joint Research Center supported UNOCHA with a rapid remote assessment after the first quake in April. Later, the UNOCHA office in Kathmandu combined indicator data from pre-disaster sources with impact data that arrived from numerous field assessments. After the second quake in June, Liew (op.cit.) created a severity index for 627 Village Development Committees (VDCs). These VDCs filled 11 out of the 14 districts on which UNOCHA was holding data. For 123 VDCs in three urban districts, including Kathmandu, food security ratings were not available; for them, the index could not be calculated.

The structure of the additively aggregated index, before population-weighting, is explained in this schematic.

Figure 14: Structure of the Nepal severity index, June 2015



Liew's trinity of impact, physical and socio-economic vulnerability was motivated by the perception that priorities should be guided by different emphases as the phases of the response progressed. Early on, in April and May, resource allocation was dictated primarily by the pattern of the impact. During the monsoon, physical vulnerability would shift the distribution of needs. As the rains abated, more attention would be paid to socio-economic vulnerability.

Liew offered two versions of the severity map – population-weighted and unweighted. He seems to have thought that they were equally important for decision makers. We will use some of his data to construct alternative models. We will compare results with his population-weighted scores (our models will all incorporate population). However, our main constructs will not follow his seasonal logic. Rather, we will build our process models with constructs that we believe are more generic and will have wider application – vulnerability (pre-existing conditions), exposure and intensity. Our measurement models for these three sub-indices will be the same across different process models. We use the Betti-Verma algorithm (implemented for Excel users in Benini and Chataigner 2014, op.cit.) to minimize redundancy among the indicators of a construct and to give higher weights to those with higher relative dispersion (i.e., more information content). With the three sub-indices, we calculate severity scores in three process models:

- An **additive** model that assumes that in determining the severity levels vulnerability, exposure and intensity compensate for each other. Weights define the substitution rates. Our weights will again be left to the Betti-Verma algorithm to determine.
- A **multiplicative** model that assumes that the three constructs compound each other. This model needs neither standardization of the sub-indices nor weighting.
- A **data-driven** model, with variable weights, calculated through Data Envelopment Analysis (DEA).

The scores from the four models (Liew's and ours) cannot be directly compared because their scales differ. However, they can be compared on a number of statistics that are indifferent to scales. We will present and discuss those statistics and evaluate the suitability of the models on them for the Nepal case study.

Measures of vulnerability, exposure and intensity

Vulnerability

We combine four measures provided by Liew: the poverty rate (headcount measure), an index of the strength of members of disadvantaged castes and ethnic groups, an index of gender inequality (based on the ratio of gender-wise literacy rates), and the proportion of residents with disabilities. Betti-Verma produced these weights:

Table 11: Indicators and weights of the vulnerability sub-index

Mean, weight and contribution to total, by item

	Mean	Weight	Contri	Share
Poverty	0.2633	0.2941	0.0774	0.3596
Caste and ethnic group	0.3730	0.1902	0.0710	0.3295
Gender inequality	0.3418	0.1721	0.0588	0.2732
Persons with disabilities	0.0237	0.3436	0.0081	0.0378
Total		1.0000	0.2154	1.0000

Chiefly the shares are of interest here. The first three indicators contribute with fairly equal and important shares to the vulnerability sub-index; disability has a small share because of relative greater redundancy with other three.

Exposure

We measure exposure as the VDC population. In theory, it would be attractive to use also residential building counts, as a second indicator. But building statistics were collected only for districts (see below).

The typical (median) population of the 627 VDCs is 3,524 persons, the mean 4,504, the minimum 414, the maximum 84,671.

For the additive model with Betti-Verma weighting, we rescale population by dividing by the maximum.

Intensity

We combine Liew's indices of building damage, earthquake-induced mortality, VDC accessibility by road (graded by vehicle type) and our ratio-level transformation of the IPC food security levels. Our transformation is based on the odds of the *ridit* (see earlier discussion on page 25 sqq.):

Table 12: Rescaling of the ordinal food insecurity scale to a ratio scale

Food insecurity				
IPC level	Ratio-rescale value	VDCs	Percent	Cum.
1	0.08	92	14.67	14.67
2	0.42	184	29.35	44.02
3	1.91	271	43.22	87.24
4	14.68	80	12.76	100
Total		627	100	

The escalation factors are high, about 1 to 8 from IPC level 3 to level 4. Basically, this is tantamount to creating an indicator with the value "1" for IPC level 4, and "0" for all others. For calculation purposes, we re-scaled the index by dividing by the maximum (=14.68).

The resulting intensity sub-index has these indicator means, weights and shares:

Table 13: Indicators and weights of the intensity sub-index

```
Mean, weight and contribution to total, by item
 -----------------------------------------------------------------
                          |   Index     Weight     Contri      Share
 -------------------------+---------------------------------------
          Building damage |  0.6715    0.0794    0.0533    0.2020
          Extra mortality |  0.2085    0.3248    0.0677    0.2566
              Road access |  0.2860    0.3005    0.0859    0.3256
          Food insecurity |  0.1929    0.2953    0.0570    0.2159
 -------------------------+---------------------------------------
                    Total |            1.0000    0.2639    1.0000
 -----------------------------------------------------------------
```

The shares are fairly well balanced; all four indicators contribute significantly.

Severity index calculation, by process model

Dispersions and correlations

The Betti-Verma algorithm is driven by the coefficients of variation of the indicators and by the correlations among them. These statistics are of a wider interest. Liew noted the negative correlation between population and his version of the severity index; Saisana (2016, op.cit.) advised against including indicators negatively correlated with others. We defend their inclusion, particularly if the process model requires them (see previous section).

Table 14: Dispersion and correlations among three sub-indices of the severity indices

Construct	Coeff.Var.	Correlations		
		Vulnerab.	Exposure	Intensity
Vulnerab.	0.27	1.00		
Exposure	1.07	-0.20	1.00	
Intensity	0.82	0.26	-0.24	1.00

We note the narrow variability of the vulnerability sub-index, compared to the wide ones of exposure and intensity. In other words, the affected VDCs were more similar to each other in terms of pre-existing conditions than in terms of population size and earthquake impact. The correlations of exposure with the others is negative, as predicted. More populous VDCs tended to be less vulnerable and less intensely affected.

Additive model: Weights by Betti-Verma

The additive model is primarily driven by the intensity of the earthquake impacts. The pre-existing conditions (vulnerability) do not contribute much to the variation of the severity.

Table 15: Severity index, additive - Relative contribution by the three sub-indices

```
Mean, weight and contribution to index, by sub-index
 -----------------------------------------------------------------
                          |    Mean    Weight    Contri     Share
 -------------------------+---------------------------------------
            Vulnerability |  0.2154    0.0855    0.0184    0.1491
                 Exposure |  0.0532    0.6467    0.0344    0.2786
                Intensity |  0.2639    0.2678    0.0707    0.5724
 -------------------------+---------------------------------------
                    Total |            1.0000    0.1235    1.0000
 -----------------------------------------------------------------
```

Multiplicative model

The multiplicative model computes the severity index as the product of the three sub-indices.

The relative contributions of the sub-indices can be approximated in a regression model of the logarithmically transformed index and sub-indices. The variance decomposition suggests strong dominance of the intensity (70 percent of the variance), at the expense of the effect of exposure (20 percent) and vulnerability (10 percent)[21].

[21] The shares of the Betti-Verma output table and the variance proportions from a regression model are not exactly comparable. Statistically minded readers might want to ask why we would not take the logic of the Betti-Verma weighting to the next level - a model with the logarithmically transformed variables. After all, the multiplicative model is additive in the magnitudes of the sub-indices. Formally, this can be done. It would result, upon re-exponentiation, in an exponentially weighted model:

$$Severity = (vulnerability)^{0.134} * (exposure)^{0.405} * (intensity)^{0.460} .$$

[fn. cont. next page]

Variable weights: Data Envelopment Analysis

The substituting or compounding effects among vulnerability, exposure and intensity may be poorly understood, and they may not be constant over their entire ranges. Thus we may want to give “the benefit of the doubt” to units that are high on one of the sub-indices, but not particularly high or even low on the others.

For example, a VDC with a large town (high on exposure), but at the fringes of the shaking zone (low intensity), may have a substantial number of poor households (medium vulnerability in terms of poverty rate). The shaking, or subsequent landslides, may have been strong enough to wipe out most of the rickety dwellings of the poor. If so, the absolute number of households in acute need is high. Conversely, a small VDC near the epicenter may have experienced total destruction and high mortality, with all its survivors found in acute need.

Analytically, the “benefit of the doubt” methodology is implemented through Data Envelopment Analysis. ACAPS offers interested readers a detailed background note and a demonstration workbook among its Web-based resources[22]. Here we limit ourselves to a visual demonstration.

In order to do so, we make a simplifying assumption that will allow us to show the functioning and some of the results to the DEA in two dimensions, i.e. in a scatterplot. We assume that the product

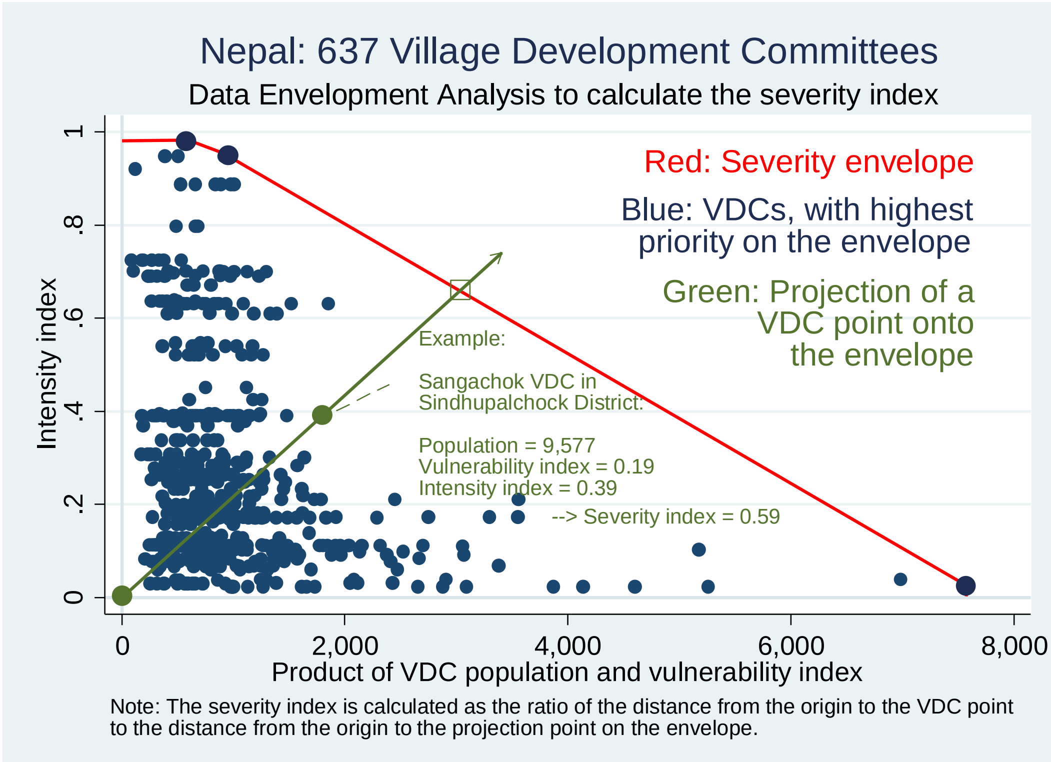
*vulnerability * exposure*

is proportionate to the number of particularly vulnerable persons in the VDC. This is against the spirit of DEA; it places a constraint on the substitution between the two constructs (the rate is one in their magnitudes – if we double vulnerability while halving exposure, the product remains the same). However, with just “intensity” and this “vulnerable-persons index” left, the model is easy to visualize.

(Betti-Verma ensures that the exponents sum to 1). However, there is no substantive theory to justify why the three constructs should be dampened, and if so, why differently.

[22] “The use of Data Envelopment Analysis to calculate priority scores in needs assessments” (2015), at: http://www.acaps.org/sites/acaps/files/resources/files/the_use_of_data_envelopment_analysis_to_calculate_priority_scores_in_needs_assessments_july_2015.pdf and http://www.acaps.org/sites/acaps/files/resources/files/data_envelopment_analysis-demo_files_july_2015.zip .

Figure 15: Data Envelopment Analysis of the severity index in Nepal



The distribution of the VDCs in this space is unusual. Most of them have a population multiplied by the vulnerability index smaller than 2,000; there are two outliers > 6,000. Most of the variation is in the intensity, which we already know from the shares reported for the additive and multiplicative models. The L-shaped distribution confirms the negative correlation between population and intensity, a fact graphically shown by Liew already. The bunching of the VDC points in horizontal streaks reflects the fact that building damage and mortality estimates were available at the district level only; the jittering up and down is chiefly due to different road access within a district (and some to the presence of highly food-insecure VDCs).

Not surprisingly then, the intensity accounts for a stunning 79 percent of the variance of the severity index, with exposure (population) claiming 11 percent, and vulnerability a mere 9 percent[23].

Comparison of four models

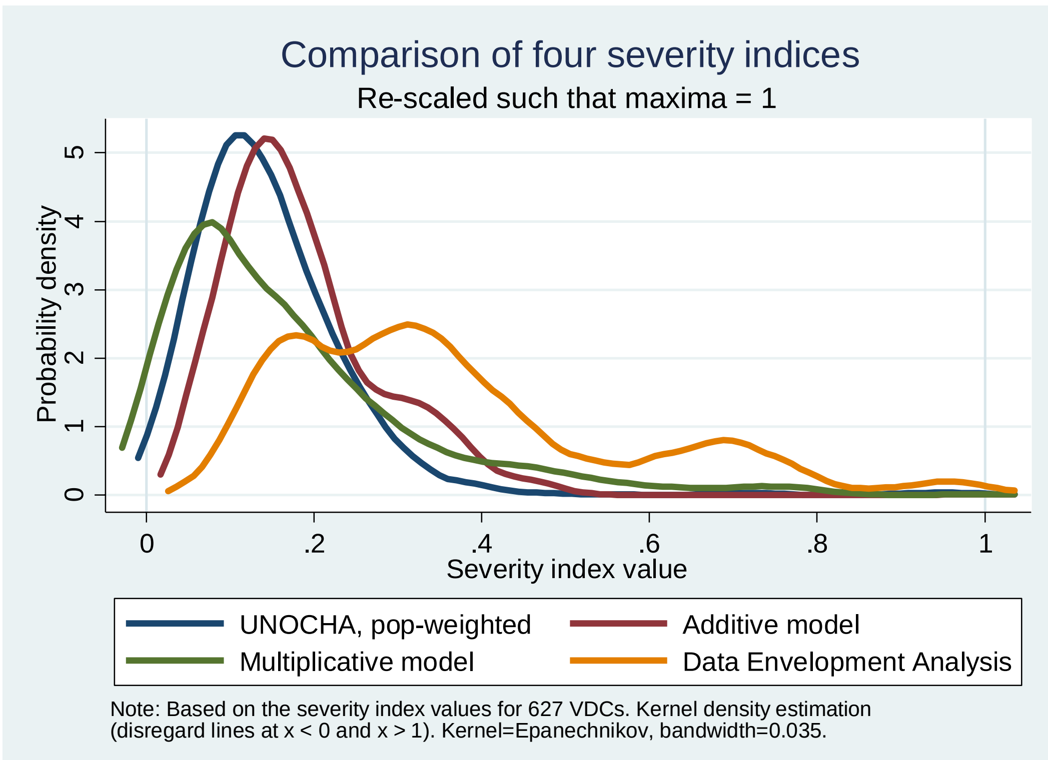
We now have severity index values from four models. The scales of these indices are not directly comparable. We make them comparable in two ways. Visually, by dividing each by its maximum value. This is not the only legitimate re-scaling; one could, for example, transform the indices such that they take the same means and variances. Analytically, we take three statistics that are robust to such transformations:

[23] Technical note: All variables of the regression model were set to their logarithms, to satisfy OLS assumptions.

- The ratio of the 90th percentile to the 50th (the median), a measure often used in concentration studies,
- The fraction of VDCs with severity values above the mean plus two standard deviations, a measure of frequency of high values from the average viewpoint
- The fraction of those with severity values above half of the maximum (i.e. on the interval (0.5, 1]), a frequency measure from the extreme-value perspective.

For the visual inspection, the next figure overlays the four distributions. These lines basically are smoothed histograms; they underplay the fine variations, but they allow bulk comparisons. What we can compare are not the means or medians, but the shapes of the curves: How strongly is the distribution skewed, how fast do the curves flatten to the right, how many distinct peaks can we make out?

Figure 16: Severity index distributions from four process models



The visual approach lets us see one difference clearly: The distribution of severity index values from the DEA model (orange line) is distinct from the other three. It is the only one with a substantial density in the upper half of the index range. One might nickname the higher densities around [0.6, 0.8] the “benefit of the doubt-bulge”, a sizeable group of VDCs that come close, but not very close, to the severity frontier, as seen in the DEA schematic (Figure 15 on page 64).

About the other three indices, the readers may have noticed that the green line (multiplicative model) in the upper range (from around 0.4) hovers above the lines for the dark blue (original model) and brown lines (additive model). But the distances are almost imperceptible; the statistics will give a better handle on the comparisons.

Table 16: Concentration statistics on four severity indices

Model	Ratio of 90th to 50th percentile	Percent VDCs with high index values	
		Above mean + 2 standard dev.	Above half of the maximum (0.5, 1.0]
UNOCHA	1.92	2.7%	1.0%
Additive	1.98	3.5%	0.2%
Multiplicative	2.84	5.3%	4.9%
DEA	2.18	4.0%	20.6%

We interpret the three statistics one by one:

- The essence of the first says: In the original UNOCHA, additive and DEA models, by doubling the median, there are only approximately 10 percent of the VDCs left with a higher severity value. In the case of the multiplicative model, even by tripling the median, there are still about 10 percent of the VDCs with higher values. In this model, we don't reach isolated outliers that early.

- The second statistic uses the reverse argument. Instead of calculating the ratio between two percentile-defined values, it defines a high cut-off and asks what proportion of VDCs are above it. Not surprisingly, it is again the multiplicative model that stands out.

- While that cut-off is based on central tendency and dispersion, the third statistic works backwards from the maximum. For the two additive models – UNOCHA's and ours - the proportions left on the upper half of the index range are vanishingly small. Another way of seeing this is to say that they produce maxima so far out that almost all other VDCs are in the lower half. Their extremes are very extreme.

It is in this third statistic that the different behaviors of the multiplicative and the DEA models are most evident. The multiplicative model caused a small, but not insignificant number of VDCs to score in this range. The DEA liberally gives the benefit of the doubt to a fifth of all VDCs; they ought to be considered high-severity by this definition.

For this sample of Nepali communities, the additive models do not seem ideal. They flatten out too soon, not leaving enough units in a reasonably defined high-severity range.

The difference between the multiplicative and DEA models is mainly philosophical. If we believe that vulnerability, exposure and intensity compound each other in equal degree – mathematically: are one-to-one substitutable in their logarithms -, then the multiplicative model should satisfy us.

This model also has the advantage that it suggests a classification of severity levels with decreasing membership, a feature that is attractive from a response-planning viewpoint. If we break the 627 VDCs into three groups of contiguous severity values such that they are maximally homogeneous, the optimal cutting points are 0.159 and 0.392. They define a low-severity group of 366 VDCs, a middling group of 200, and a high-severity one of 61[24]. In other words: Pay attention first and foremost to about one tenth of the affected units.

In a sign of humility, we may admit that we don't know the relationships among the key constructs that well. Our process models may not lead us to adequate descriptions of the suffering of those affected by crises and disasters. To some communities that the additive and multiplicative models, in our hubris of assumed knowledge, relegate to perceptions of low severity, we ought to give the benefit of the doubt. They may truly be hit as hard as those that appear in our top category of severity. If we have that attitude, the DEA model plays it safer for us and for them.

For completeness, here are the results of the same homogeneity maximizing procedure applied to the DEA index values: Cut the continuum at 0.271 and 0.534. This forms a low-severity group of 230 VDCs, a medium one of 277, and a high one of 120. In other words: Pay attention first and foremost to about one fifth of the affected units.

This case-study has been about Nepal and about its data collected on communities affected by earthquakes. It goes without saying that the distributional characteristics of additive, multiplicative and DEA models may turn out very different when they are calculated for other regions or from other indicators. Our demonstration is merely didactic, to show the differences in outcomes, depending on the models chosen. It is not prescriptive in an immediate technical sense. In any context, the choice of additive, multiplicative vs. DEA models (or any others that seem worth exploring for severity measures) must be driven by our conceptual understandings. Only when these are sufficiently settled, can we proceed to define the quantitative relationships.

Outlook

In the structure of this note, the major switch is between severity measures that combine sector-based information and those that rely on indicators most or all of which are not tied to sectors. Logically, syntheses between the forms are possible. Practically, institutional factors and the information landscape will privilege one over the other, not so much as a choice of the preferred option than in the response to situational feasibility.

Whichever avenue the effort to measure severity takes, in the end the synthetic measure – inter-sectoral or multi-indicator – has to be justified as representing "severity". If the affected unit A (e.g., a village community) has a higher severity score than B, independent observers, given the information that went into the scores, should be able to agree that A is in a more severe condition than B. That information consists not only of the data, but also of the rationales and methods whereby the data were fused.

[24] Using a one-dimensional clustering procedure that minimizes the sum of the within-cluster sums of squared deviations from cluster means (Fisher 1958). The cut-offs are from the transformed index (max. = 1).

The rationales and methods discussed in the preceding sections leave a number of areas open. In these more work and better insights are needed in order to ensure that severity measurement remains practical and legitimate. This list is far from complete:

- Validation: Contrary to instruments that measure severity in other domains (particularly health), humanitarian severity measures are not validated. This is systemic: if severity determines the humanitarian response, the response will avert the outcomes that the severity measures predicted (in this sense severity measure are self-destroying prophecies). Validations, therefore, will come not only from observed outcomes among the affected, but even more so from changes in attention and engagement among the responders. For this, new models and data will be needed.

- Uncertainty: Currently, as far as we know from shared severity assessments, the measurements have all been deterministic. The uncertainty resulting from measurement error, from the lack of validated process models and, where it is done, from sampling is rarely noted, much less estimated. It should be incorporated. Elicitation methods that prompt experts and key informants to provide for each of their estimates the most plausible value (or probability for a binary indicator) as well as minimum and maximum provide a starting point for sensitizing the humanitarian information management community. Simulation methods are equally helpful. Sensitivity or robustness analyses are recommended, but are rarely undertaken for lack of time and expertise (For an example, from Typhoon Haiyan in the Philippines, see Benini and Chataigner 2014:60-63).

- Ordinal variables: The aggregation of ordinal variables remains a challenge on which academic researchers have done considerable work. The Borda count (see page 52 sqq.) dealing with rankings is on firm ground. Methods combining ratings such as severity levels are less well known; our proposal for a data-driven approach (the odds of the ridit) is more of a deliberate provocation than a solution that will pass methodological muster. Better practices are likely once some promising algorithms are translated into something like Excel macros.

- Heterogeneous affected units: The units that assessments compare on severity are highly variable. Affected communities include far-flung rural tracts with scarce populations side by side with towns and cities that compress tens or hundreds of thousands into small areas. The negative correlation between intensity and exposure results from it; we found it in Nepal, but it will likely be observed in other theaters as well. The multi-resolution challenge is another facet of the heterogeneity. The social geography, measurement and expert judgment literatures should be scoured for ways of better dealing with it.

- Much of our discussion is limited to the “one observation per affected unit” format. The Big Data revolution will explode this format. Apart from the greater speed and volumes imposed on collection, analysis and sharing (of what? of real-time results? - there may no longer be the traditional final report), aggregation challenges will re-surface in new guises. Selection bias will likely be more of a plague on the measurement side; model selection techniques more likely a boon on the process modeling side. Some of the current methodological insights will be scaled up to Big Data, and others will be lost or irrelevant and replaced.

While so many things are in flux, it is important to remember that severity measures have a single function: to focus attention on affected persons whose condition is worse than that of others, and to thereby inform the humanitarian response. The process models of severity must be able to capture these differences, the measurement models must correctly quantify the constructs, and the data collection and management must supply the quantities needed. Big Data, Small Data, whatever data - there is a hierarchy that runs from function/purpose down to process to measurement to basal perceptions, and back upwards from the data bit to, ultimately, informed decision-making. Improvements in severity measurement happen when we run up and down this staircase, armed with big brooms and tiny brushes.

Recommended practices

Determine the major type of the severity measure early on

During the time when the objectives of the assessment are being clarified, determine whether the production of consistent, universal severity measurements throughout the affected region is feasible. Considering the humanitarian history and the institutional configuration, determine whether the needed information will arrive primarily from sector-specific collections or from indicators and other material not couched in sectoral definitions.

If the measure is going to be sector-based:

If sector-specific, estimate the chances that the sectors will supply reasonable estimates of persons in need and of severity levels, or whether their assumptions call for fixed or constrained estimates. Conduct preparatory events with sector coordinators and their information managers to advance common understandings of need, severity and their measures, or to make irreconcilable differences explicit if any persist. Offer formats in which sectors would be willing to report estimates of persons in need by different levels of neediness, ideally with plausible, minimum and maximum values. Persons in acute need, in moderate need, not in need for humanitarian assistance form one of several possible category sets. A more radical approach would ask the sectors to estimate persons with catastrophic, critical, severe or minor/major problems, in a scheme with four levels, accompanied by nutshell evidence for every estimate (see Table 7 on page 45).

Determine the aggregation strategy to combine the various sectoral estimates. For lack of a better formula, use the maximum of the sectoral PIN estimates in a given unit as the unit's combined PIN. Be aware that, if estimates are obtained for PIN at different levels of neediness, the use of the maximum formula may cause paradoxes – across the neediness levels totals may exceed the population and will then need a proportionate adjustment. If the sectors indeed supply estimates with ranges (minimum, most plausible value, maximum), draw multiple random estimates (e.g., from triangular distributions); compute the above maximum for each replication and average. If the number of replications is sufficient (≥ 20), then calculate also 90-percent confidence intervals.

Limit the sectors whose figures are to be combined to those that address needs of a common nature. Thus, sectors that respond to needs for short-term material and services (e.g., shelter, health, food, NFI) may form one group; sectors working on longer-term rehabilitation (e.g., Education) a different group; Protection one in this own right. Review the estimates for these groups, separately applying the maximum formula, side by side and determine whether a combined PIN estimate for all makes sense, or loses so much discrimination that separate estimates serve responders better.

If the measure is to be based on indicators:

If the severity measure is to be built with non-sector-specific indicators, build your process model considering the type or types of crisis and disasters. Our basic model works with three constructs – vulnerability, exposure, intensity -, and the quantitative operations connecting them. If the model is enriched by other constructs – such as humanitarian access -, the additional relationships too need to be mathematically defined (e.g., in a way that avoids division by zero).

For every construct, build a measurement model. If there are multiple indicators to choose from, include enough so that missing values (if they are few) can be imputed. Determine whether the concept behind the construct is cause or consequence of the indicators. If the construct causes the indicators, expect them to be correlated, and take the absence of correlations among the indicators as a warning of problems. If it is the consequence, do not expect correlations, but ensure that indicators cover all aspects defining the concepts (see example of poor families living in low plains in flood risk areas). Determine whether the construct should maximize the diversity in the information carried by its indicators, or maximize commonality. In the first case, apply the Betti-Verma double-weighting rule (see detailed explanations in Benini and Chataigner 2014, op.cit.), in the second estimate calculate factor scores (call for statistical help).

Keep models simple. Any process model for the severity measure and any measurement model for a constituent construct should find enough space on the back of an envelope. If severity measures on the same units or in the same or a similar context were produced before (or are currently being designed by others), try to ensure that yours uses measures that are comparable, as long as their formats do not interfere with your essential purposes.

In the analysis of indicator-based measures, inspect the distributions of the indicators and then of the constructs over their entire ranges. Are gaps and outliers responsible for unexpected distributions at the next higher level? Are they posing challenges to reliability and/or validity? Consider corrections in the measurement models of the constructs.

The assumptions that you made in your process model should have led to the choice of one and only one major model (additive, multiplicative, or DEA); so most likely, you will not need to do model comparisons at that level. If you do, standardize the severity indices from several models in a way useful for comparison, such as dividing each of them by its maximum, and compute scale-free diagnostic statistics as shown in the Nepal case study.

Present the severity measure together with its major constructs

In the final listing of severity-scored units, users may appreciate a visual format that shows them which of the three constructs have contributed to the severity index value for a given unit in what proportions. There are several ways of doing this without information overload. A quick way, though limited to a relative intuition, is to color each of the three columns that hold the construct values using color scales in Excel's conditional formatting. Tables may be sorted on the severity index or on any of its components, but never replace the ratio-level severity measure with its ranks. A four-panel map with choropleth representations of the severity index and of the three constructs lets users see how they spatially diverge (see the example in the Summary). Such differentiated representations may be helpful in adapting response strategies to sets of affected units whose severe conditions are driven by different principal factors.

Table 17: Visualization of the relative importance of the constructs, unit by unit

Area	Vulnerability (Poverty rate)	Exposure (Population)	Intensity (Fraction destroyed buildings)	Severity index
District 4	0.45	9,000	0.76	1.00
District 5	0.40	11,000	0.67	0.96
District 6	0.35	13,000	0.53	0.78
District 2	0.55	5,000	0.74	0.66
District 3	0.50	7,000	0.58	0.66
District 1	0.60	3,000	0.50	0.29

Note: This is an extremely simple scenario, with just six areas, and with each of the major constructs measured by just one indicator. The districts are sorted descendingly on their severity scores. The values of the constructs are visible, but their relative importance, within their respective columns and in each row, is rapidly assimilated through the color gradations from conditional formatting. The formatting is done separately in each column.

Appendix: Example of a severity scale

Scale created by a sector coordination unit: WASH sector for the Syria 2016 HNO

	No need of external assistance		Need of humanitarian assistance		Acute and immediate need of humanitarian assistance		
	0	1	2	3	4	5	6
TOPICS	No problem	Minor Problem	Moderate problem	Major Problem	Severe Problem	Critical Problem	Catastrophic Problem
Magnitude of problems in terms of population number	Normal situation. No displaced people and living condition is normal.	At least 80% of the population in the sub district have access to improved water source	At least 40% of the population in this sub district don't have access to improved water source that led to an increase of skin diseases.	At least 50% of the population don't have access to improved water source that led to an increase of skin diseases and diarrhoea.	At least 70% of the population don't have access to improved water source that could lead to severe effect on health (skin diseases and acute water diarrhoea).	At least 70% of the population don't have access to improved water source that could lead to irreversible effect on health situation (High Acute water diarrhoea rate, death etc.)	At least 70% of the population don't have access to improved water source and there is visible impact on health such as death of high acute water diarrhoea reported ect…
Coping mechanisms	Population is living under normal conditions.	People with no access to improved water sources are getting their water for personal consumption from at least one of the following source: - Damaged water infrastructure - Traditional wells	People with no access to improved water sources are getting their water for personal consumption from at least one of the following source: - Damaged water infrastructure - Traditional wells And have to reduce the amount water for personal consumption (less than 15l per person and per day)	People with no access to improved water sources are getting their water for personal consumption from at least one of the following source: - Damaged water infrastructure - Traditional wells - Water trucking And have to reduce the amount water for personal consumption (less than 10l per person and per day)	People with no access to improved water sources are getting their water for personal consumption from at least one of the following source: - Damaged water infrastructure - Unprotected spring - Traditional wells - Unprotected spring - Water trucking And have to reduce the amount water for personal consumption (less than 7,5l per person and per day)	People with no access to improved water sources are getting their water for personal consumption from at least one of the following source: - River - Pond - Water trucking And have to reduce the amount water for personal consumption (less than 5l per person and per day)	People with no access to improved water sources are getting their drinking water from: - River, - Pond And have to reduce the amount water for personal consumption (less than 5l per person and per day).
Water Access	Water prices and access are following normal seasonal trends	Only minor constraint impeding regular and easy access	Moderate constraints (above normal prices, small acceptable delays due to roads, distance, checkpoint, etc.) impeding regular and easy access	Major constraints (high prices, large delays due to roads, large distance, waiting at checkpoint, insecurity on the road, etc.) impeding regular and easy access	Severe constraints (very high prices, waiting time, physical risks, low purchasing power) impeding regular and easy access	Critical constraints (very high prices, waiting time, life threatening risks, very low purchasing power) impeding regular and easy access	No access
Water Availability	Water available in the area following normal seasonal trends	Light/minor shortages of water	Moderate shortages of water	Major shortages of water	Severe shortages of water	Critical shortages of water	Safe water is not available

References

ACAPS and et.al.; (2011). Report of findings of multi-sector, inter-agency assessment on humanitarian needs in the districts of Satkhira, Khulna and Jessore. Field research conducted 8th-12th September 2011 [Main editor: Sandie Walton-Ellery], ACAPS and several national and international organizations.

Adelman, H. and A. Suhrke (1996). "Early warning and response: Why the international community failed to prevent the genocide." Disasters **20**(4): 295-304.

Beccari, B. (2016). "A comparative analysis of disaster risk, vulnerability and resilience composite indicators [published 14 March 2016]." PLoS Current Disasters, Retrieved 28 July 2016, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4807925/.

Benini, A. (2015a). Moderate need, acute need. Valid categories for humanitarian needs assessment? Evidence from a recent needs assessment in Syria [26 March 2015]. Geneva and Beirut, ACAPS and MapAction.

Benini, A. (2015b). The use of Data Envelopment Analysis to calculate priority scores in needs assessments [8 July 2015]. Geneva, ACAPS.

Benini, A. and P. Chataigner (2014). Composite measures of local disaster impact-Lessons from Typhoon Yolanda, Philippines. Geneva, Assessment Capacity Project (ACAPS).

Benini, A., P. Chataigner, N. Noumri, L. Tax and M. Wilkins. (2016). "Information gaps in multiple needs assessments in disaster and conflict areas. With guidelines for their measurement, using the Nepal 2015 earthquake assessment registry." Geneva, Assessment Capacity Project (ACAPS). Retrieved 23 May 2016, from http://www.acaps.org/sites/acaps/files/resources/files/info_gaps.pdf.

Betti, G., B. Cheli, A. Lemmi and V. Verma (2005). On the construction of fuzzy measures for the analysis of poverty and social exclusion. INTERNATIONAL CONFERENCE TO HONOUR TWO EMINENT SOCIAL SCIENTISTS C GINI AND MO LORENZ. Annals... University of Siena.

Blumenthal, D., E. Malphrus and J. M. McGinnis (2015). Vital signs: core metrics for health and health care progress. Washington DC, National Academies Press.

Borda, J. C. (1781). "Mémoire sur les élections au scrutin." Histoire de l'Académie royale des sciences **2**: 85.

Borges, M. A. and B. K. Sawyers (1974). "Common verbal quantifiers: Usage and interpretation." Journal of Experimental Psychology **102**(2): 335.

Bross, I. D. (1958). "How to use ridit analysis." Biometrics: 18-38.

Bureau of Labor Statistics (US) (1916). Hazardous occupations. U. S. B. o. L. Statistics. Washington DC, U.S. Government Printing Office. Industrial Accidents and Hygiene Series, Issues 9 - 16.

Charman, T. and K. Gotham (2013). "Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders–lessons from research and practise." Child and adolescent mental health **18**(1): 52-63.

De Boer, J. (1990). "Definition and classification of disasters: introduction of a disaster severity scale." The Journal of emergency medicine **8**(5): 591-595.

De Groeve, T., K. Poljansek and L. Vernaccini (2015). Index for Risk Management - INFORM: Concept and Methodology [Version 2016]. Luxembourg, Joint Research Center of the European Commission (JRC).

Elcheroth, G., S. Penic, R. Fasel, F. Giudici, S. Glaeser, D. Joye, J.-M. Le Goff, D. Morselli and D. Spini (2013). "Spatially weighted context data and their application to collective war experiences." Sociological Methodology **43**(1): 364-411.

Eriksson, A., Y. K. Ohlsén, R. Garfield and J. von Schreeb (2015). "Who Is Worst Off? Developing a Severity-scoring Model of Complex Emergency Affected Countries in Order to Ensure Needs Based Funding." PLoS Currents **7**: ecurrents.dis.8e7fb95c97df19c95a99ba56584d56586aa56582c56559.

Fisher, W. D. (1958). "On Grouping for Maximum Homogeneity." Journal of the American Statistical Association **53**(284): 789-798.

Forbus, K. D. (1996). Qualitative Reasoning. Draft chapter for the CRC Handbook of Computer Science [4 June, 1996]. Evanston, IL, Qualitative Reasoning Group. The Institute for the Learning Sciences, Northwestern University.

Fornell, C. and F. L. Bookstein (1982). "Two structural equation models: LISREL and PLS applied to consumer exit-voice theory." Journal of Marketing research: 440-452.

Harff, B. and T. R. Gurr (1998). "Systematic early warning of humanitarian emergencies." Journal of Peace Research **35**(5): 551-579.

Hasenfeld, Y. (1972). "People Processing Organizations: An Exchange Approach." American Sociological Review **37**(3): 256-263.

Humanitarian Liaison Group (2014). MSNA. Syria Multi-Sectoral Needs Assessment [October 2014], Prepared by OCHA, REACH and SNAP on behalf of the Humanitarian Liaison Group based in Turkey.

IPC Global Partners (2008). Integrated Food Security Phase Classification Technical Manual. Version 1.1. Rome, FAO.

IPC Global Partners (2012). Integrated Food Security Phase Classification Technical Manual. Version 2.0. Evidence and Standards for Better Food Security Decision. Rome, FAO.

Jarvis, C. B., S. B. MacKenzie and P. M. Podsakoff (2003). "A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research." Journal of Consumer Research **30**(2): 199-218.

Jenkins, S. P. (2008). Estimation and interpretation of measures of inequality, poverty, and social welfare using Stata. North American Stata Users' Group Meetings 2006, Stata Users Group.

JRC, UNOCHA and Acaps (2015). Toward the development of a global severity index. Ispra, Italy and Geneva, Switzerland, Joint Research Center (JRC), United Nations Office for the Coordination of Humanitarian Assistance (UNOCHA), and Assessment Capacity Project (Acaps).

Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. P. Slovic and A. T. (eds.). Cambridge, UK, Cambridge University Press.

Kuipers, B. and D. Berleant (1988). Using Incomplete Quantitative Knowledge In Qualitative Reasoning. AAAI-88 Proceedings. **88:** 324-329.

Liew, A. (2015). "Concept note: Severity Index [16 July 2015]." Kathmandu, Nepal Earthquake Assessment Unit. Retrieved 28 July 2015, from https://www.humanitarianresponse.info/en/operations/nepal/document/150716-severity-index.

Limbu, M., L. Wanyagi, B. Ondiek, B. Munsch and K. Kiilu (2015). "Kenya Inter-agency Rapid Assessment Mechanism (KIRA): A Bottom-up Humanitarian Innovation from Africa." Procedia Engineering **107**: 59-72.

Mosteller, F. and J. W. Tukey (1977). "Data analysis and regression: a second course in statistics." Addison-Wesley Series in Behavioral Science: Quantitative Methods.

Munda, G. and M. Nardo (2009). "Noncompensatory/nonlinear composite indicators for ranking countries: a defensible setting." Applied Economics **41**(12): 1513-1523.

National Research Council (U.S.) (1945). Fasciculus on Chemical Warfare Medicine, Volume I, Eye. Washington DC, National Research Council (U.S.).

Porter, T. M. (1995). Trust in Numbers. The Pursuit of Objectivity in Science and Public Life. Princeton, Princeton University Press.

Robinson, W. S. (1950). Ecological Correlation and the Behavior of Individuals, American Sociological Review: 351-357.

Rutstein, D. D., R. J. Mullan, T. M. Frazier, W. E. Halperin, J. M. Melius and J. P. Sestito (1984). "Sentinel health events (occupational): a basis for physician recognition and public health surveillance." Archives of Environmental Health: An International Journal **39**(3): 159-168.

Saisana, M. (2015). Composite Indicators and Scoreboards. Workshop on Crisis Severity Index, 21 April 2016. Ispra, Joint Research Centre.

Samejima, F. (1969). "Estimation of latent ability using a response pattern of graded scores." Psychometrika monograph supplement.

Sen, A. K. and J. Drèze (1999). The Amartya Sen and Jean Drèze omnibus : comprising poverty and famines, hunger and public action, India: economic development and social opportunity. New Delhi ; New York, Oxford University Press.

Slovic, P. (2000). Perception of Risk. London, Earthscan.

UNOCHA (2015). 2016 Humanitarian Needs Overview. Syrian Arab Republic [version 18 October 2015]. New York and Geneva, United Nations Office for the Coordination of Humanitarian Affairs, in collaboration with humanitarian partners.

Velleman, P. F. and L. Wilkinson (1993). "Nominal, ordinal, interval, and ratio typologies are misleading." The American Statistician **47**(1): 65-72.

Wikipedia. (2011a). "Borda count." Retrieved 28 March 2011, from http://en.wikipedia.org/wiki/Borda_count.

Wikipedia. (2011b). "Likert scale." Retrieved 28 October 2011, from http://en.wikipedia.org/wiki/Likert_scale.

Wikipedia. (2014). "Anscombe transform." Retrieved 19 February 2016, from https://en.wikipedia.org/wiki/Anscombe_transform.

Wikipedia. (2016a). "Anchoring." Retrieved 15 March 2016, from https://en.wikipedia.org/wiki/Anchoring.

Wikipedia. (2016b). "Odds." Retrieved 22 March 2016, from https://en.wikipedia.org/wiki/Odds.

Wikipedia. (2016c). "Polytomous Rasch model." Retrieved 24 May 2016, from https://en.wikipedia.org/wiki/Polytomous_Rasch_model.

Wu, C.-H. (2007). On the application of grey relational analysis and RIDIT analysis to Likert scale surveys. International Mathematical Forum.

Zheng, X. and S. Rabe-Hesketh (2007). "Estimating parameters of dichotomous and ordinal item response models with gllamm." Stata Journal **7**(3): 313-333.