Aldo Benini

A note for ACAPS

Severity and priority -

Their measurement in rapid needs assessments

---

8 August 2013

Table of Contents

Tables and Figures

Summary

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

The Second Joint Rapid Assessment of Northern Syria (J-RANS II), published in May 2013 (AWG 2013b), presents an opportunity to clarify the conceptual relationship between those two measures. The assessment employed severity scales in five sectors (public health, food security, nutrition, shelter, water and sanitation) as well as a priority scale for seven sectors (those five plus education and protection).

This note investigates the construction and functioning of those scales and the correlations between severity and priority. For contrast, we discuss the same aspects of an assessment earlier carried out in Yemen, whose data architecture and link between the two measures were significantly different from the J-RANS II. Both assessments, however, struggled with the fact that their severity scales differentiated poorly. This shortcoming motivated us to generate an artificial dataset to simulate what associations can realistically be expected between severity and priority measures.

We make seven practical recommendations:

1. Continue using severity as well as priority measures needs
2. Revise the severity scale (and how)
3. Make a minor change in the priority scale
4. Experiment with a second severity measure
5. Continue expressing severity with population figures
6. Elicit the "serious problems" (= "key issues") differently
7. Where the enumerators are well educated, make greater use of them.

With several assessments done, ACAPS has the experience and knowledge for cumulative learning. This includes how measures of unmet needs work or don't work. Generic measurement tools can be distilled and canned, and then shipped to new theaters and used "with some local assembly, as required."

Introduction

This note speaks to the measurement of needs in rapid needs assessments, specifically to ways of expressing the severity and priority of different needs. The note was occasioned by a review of a dataset from a recent assessment in northern Syria (AWG 2013b). It discusses also the approach to severity and priority chosen in an earlier assessment in Yemen. Differences in concepts and data architectures between the two assessments were substantial, yet the challenges of valid measurement and correct analysis were similarly daunting. The purpose is to derive lessons that can be applied to the design of future needs assessments.

There are two key questions:

- One is about valid measurement. Here the challenges appear to be taller when we try to establish severity. The degree of severity is a property inherent of every unmet need. Statistically, it is independent of the severity of other needs, but this benefit is more than offset by the difficulty to find a common language across need domains. By contrast, priorities arise from comparisons among two or more needs. Although this creates statistical dependence, with its associated problems, in terms of common-sense approaches to needs, priority is the easier to understand of the two concepts. "Food, in our current desperate situation, is more important than education", a respondent might say, expecting to be understood by others without further translation.

- The second question is whether, by using severity and priority measures in the same assessment instrument, they somehow corroborate each other's validity. This is relevant at the aggregate as well as at the unit level. If, in the sample average, both severity and priority measures identify the same highest needs, we give more credibility to these findings than if we had one measure only, or if the two measures contradicted each other. Similarly, a strong positive correlation between severity and priority of a given need over the assessed areas strengthens trust in the assessment.

The two questions do not receive the same degree of support from social science. Severity is well understood for specific domains, as in "severe weather" or "severe financial crisis". But there is no uniform relationship between the concept of severity and human needs across all major domains. A severe food crisis has people starving in the short or medium term; a severe educational crisis impoverishes communities in the long run. Moreover, people affected by crises may delineate need domains differently from the way humanitarian agency and assessment teams define them. Thus the latter, for reasons of mandate, may conflate shelter and non-food items in a common sector. People on the ground see no need to lump them in their survival plans. These anyway they formulate in terms of events, budgets and networks rather than "sectors": *"Do I still keep my credit with the corner store selling us groceries, baby formula and kerosene until my husband brings home his next salary if he will get paid at all?"* It is thus not surprising that the tools for measuring the severity of needs in rapid assessments have evolved more from

best-practice reflection, and also by mere improvisation, within specific communities of practice, rather than from universal standards.

The relationship between severity and priority measures, however, has been extensively investigated, even if this took place under different conceptual titles and in academic circles distant from humanitarian agencies. By and large, the difference between the two types of measures boils down to that between *rating and ranking*.

- Rating assigns to an object a member of an ordered set of elements. Such sets come in the shape of scales, ladders, or verbal importance statements.
- Ranking creates an order among the objects themselves, defined by a preference or dominance relationship.

Rating systems allow the same position in the order to be filled with any number 0 to N of the rated objects. Ranking systems (usually) minimize ties.

The pros and cons of either system have been studied for a long time, in public opinion research and particularly in international values studies (Alwin and Krosnick 1985; Maio, Roese et al. 1996; Klein, Dülmer et al. 2004). The specific question that interests us the most - the synergy from concurrent use of rating and ranking -, however, has rarely been directly approached (De Chiusole and Stefanutti 2011; Roszkowski and Spreat 2012)[1].

This imbalance between the difficulty to penetrate "severity" in needs measurement on one side and the fairly well tested rating and ranking tools on the other makes it necessary at this point that we improvise some conceptual assumptions. We assume that

- "Severity" expresses the degree of *unmet needs*. It is thus related to shortages and deficits, as opposed to fulfillment and wellbeing. Severity in different domains can be expressed in very different measures, such as the price of bread for food, and the number of hospitals no longer operational for health care. However, if distinct needs are to be compared regarding their severity, a common grammar or procedure for elaborating such comparisons is needed. For example, shortages in different domains could be evaluated for their likely impact on mortality. Humanitarian assessments trying to establish the severity of needs across sectors are majorly concerned with designing and validating such procedures.
- Second, in every sector unmet needs may be thought of as an *underlying continuum* that is only indirectly observed. The observations may be in continuous measures for particular needs, such as the price of bread, but the quest for measures applicable in multiple sectors makes it likely that assessments must work with *discrete, categorical* definitions[2]. The necessity to express degrees of unmet needs then typically produces *ordinal* measures.

[1] The first of these two articles was accessible only in its abstracts, the second in its first two pages.

[2] For practical reasons. Theoretically, it is possible to define unmet needs in terms of waiting times, which are continuous measures, e.g. "how many days for the average resident to consume 10 liters of drinking water, 3000 calories of food, for patients with appendicitis to get surgery, etc.".

- Third, under certain conditions, statistics of ordinal measures attain *interval-level* quality, thus leading to more informative comparisons. This is feasible in different situations. We can think of three; more are conceivable: if several concurrent severity measures are taken of the same need; if a model of the structure of several underlying continuous needs exists; or if the priorities among needs are generated in analogy to an election system. These call for different statistical analyses. In the first case, we consider Likert scales (Wikipedia 2011b), in the second some form of principal components (Kolenikov and Angeles 2004), in the third a measure known as the Borda count (Wikipedia 2011a).
- Fourth, assessment teams and assessed communities do not necessarily think in terms of the same need areas (see above, shelter and NFI). The measurement of sector-specific needs may thus have to be broken down into sets of more specific measures, or at least extended by supplementary measures. These operationalizations can take different conceptual routes[3], with categorizations and re-categorizations performed by different actors at different stages of the data collection and analysis process.
- And finally, as a result of the above, we anticipate different data architectures from assessment to assessment, each with its particular data management needs before analysis can start.

The minimal ingredients for a needs-measuring system, in rapid assessment perspective, thus are: "sector" (or "domain", or any term that bundles related needs), "problem" (or "issue" or any term that expresses specific manifestations of deficits or of their consequences), rating (notably on severity scales), ranking (in terms of priority), comparison (on the basis of whatever statistics are appropriate, given the measures). One might add population (or any other important characteristic of the units having unmet needs that would call for weighting) and the operation of measuring itself, particularly the extent and impact of measurement error. These too will be discussed, if selectively.

Returning to the formal side of severity and priority, this table summarizes the pros and cons of ratings and rankings, as they have been noted repeatedly by researchers. We borrow heavily from Roszkowsky and Spreat (2012: 59-60).

[3] As they indeed have in the case of the Syria and Yemen assessments. Thus, the former prompted key informants to directly rate and rank needs sectorwise. The severity rating within each sector was followed by some form of prioritizing problems from predefined lists, but this operation had no direct impact on the severity or priority measures. In Yemen, by contrast, communities would rate "issues" within each sector, which were assigned to categories by the assessment teams. In addition, the teams would give a "synthesis" rating about all the issues that a community raised in a given sector, and communities would designate three sectors as their first, second, and third priorities. We revert to this in greater detail in later sections.

Table 1: Strengths and weaknesses of rating and ranking

Rating		Ranking	
Pro	Contra	Pro	Contra
Equivalence between items is validly expressed by equal levels on the scale.	Low differentiation (= overuse of the same score across items) due to social desirability or extreme response bias	Forces differentiation	Differences between items may be artificial if subjects view them as equivalent.
Degrees of difference can be expressed on the scale			Rank differences do not express degrees of difference.
Less time-consuming in interviews.	Attention often superficial, unfocused.	Respondents pay more attention.	More time-consuming in interviews.
			First and last items on a list tend to be over-ranked, middle items under-ranked.
	List dependency: Although each item is formally independent, the number of items influences how each is rated (learning effects during the interview).		List dependency: Whether subjects choose X > Y or Y > X, depends on how many other items are to be ranked. Lower ranks very unreliable. If many items are ranked, all unreliable.
Item ratings are statistically independent.			Statistical dependence of ranks makes certain analyses problematic.
Both subjects and items can be scored (e.g. by medians or frequencies of values in ranges of interest)	Ordinal data: legitimate stats limited to frequencies, medians, minima and maxima	If subjects understand limited choice situation, Borda counts on interval level (ratio only if meaningful zero point)	No total score for subjects (since rank sum equal for all)

The academic debate over the ultimate superiority of ratings vs. rankings is ongoing. Outside academia both continue to be practiced widely, often with drastic consequences for entire institutional sectors (sports, US college rankings). As mentioned, the combined use of rating and ranking in the same instrument has not been widely studied. Of note are the conclusions of Roszkowski et al., op.cit., that *"typically, at the aggregate level,*

rankings and ratings lead to the same conclusion" whereas [at the individual level] *"the lack of differentiation in ratings is one reason producing inconsistencies between ratings and rankings"*.

When we translate these insights back to the world of rapid needs assessments, we come to expect significant, yet modest correlations between the (rated) severity and the (ranked) priority of a given need. This for a number of reasons:

- Both measures express the underlying real needs in discrete manner, the severity score by categorization, the priority score as a rank variable.
- The severity score has a limited range, truncating extremely low and high needs to its endpoint categories.
- The priority score loses some information by assigning a zero score to more than one sector (those not considered a priority when priority options are limited). This amounts to a "modified Borda score" (Wikipedia 2011a), with the consequence that the sector-wise Borda counts have no meaningful zero points.
- The scoring of both severity and priority is subject to measurement error, the extent, direction, and correlation of which mostly remain unobserved.

These limitations make our two main questions of interest all the more urgent:

1. How should severity scales be designed?
2. In relation between severity and priority,
    a. over the assessed individual locations (e.g. sub-districts, camps), given the sector, what is the expected correlation between the severity and priority scores?
    b. for the entire sample, do both measures identify the same sectors as those with the highest needs?

We will now try to illuminate them with select results from the Syria and Yemen assessments.

The note proceeds as follows: After a minimum of background information on the assessments in Syria, and on the data architecture of the J-RANS II, we analyze the use of the severity scale made in this assessment. We consider minor adaptations, and leave major alternatives to a later section. We observe how it worked together with the priority scale at the unit as well as at the aggregate levels. We offer some comments on how "problems" within sectors were elicited.

A chapter on an earlier assessment in Yemen follows. We proceed similarly, with an emphasis on how sector prioritization and key issue rating can be combined to produce a list of priorities, not at the sector, but at the key issue level. While difficult, this alternative to the J-RANS format is still noteworthy.

We then move away from the specific country context to develop general considerations for the measurement of severity and priority. We present possible alternatives to the

severity scales as used before. We recommend one as a complement, not a replacement. We make six more practical recommendations. We conclude with the hope that cumulative learning is possible also with regards to the instruments of measuring the severity and priority of unmet needs.

[Sidebar:] Rating and rank aggregation

The process of combining ratings and rankings into summary indices (such as of the severity and priority of needs by sectors over all assessed locations) is known as aggregation. There is a substantial literature on rank aggregation, with roughly 2,500 works returned in a Google Scholar search. The rating aggregation literature is smaller, with only 300 or so hits. Rank aggregation research has soared in recent years, driven by the interest in refining page ranking in Web searches.

Although the logic of aggregation algorithms is fascinating, the literature is not very relevant for the purposes of rapid needs assessments. On the ranking side, the Borda count is both sufficient and convenient. The severity scores result from ratings; aggregation algorithms are mathematically demanding and, to our knowledge, none are available in Excel. Moreover, the complications of aggregating rankings from purposive samples with any of those methods are unknown (not even mentioned in the literature). Median scores and population figures, broken down by regions, for certain severity levels ("at risk") suffice for most purposes.

For readers willing to face the mathematics, the book *"Who's #1? - The Science of Rating and Ranking"* (Langville and Meyer 2012) provides a readable, yet solid introduction. However, most of the examples are taken from the world of US American ballgame sports. The ratings and rankings in this domain are derived from count variables with virtually no measurement error - points scored or lost in the games. These methods are of little interest in turbulent environments that reduce humanitarian assessments to weak metrics and to significant measurement error.

Of some potential interest is a method that the authors demonstrate for rating objects on the basis of *incomplete ratings* - such as when, hypothetically, key informants in different localities respond to severity questions for variable subsets of sectors. Langville and Meyer developed a matrix algebra method and illustrated it with simulated data on ratings given for National Science Foundation funding applications. Different reviewers rated different, but overlapping subsets of applications (pages 179-181). Conceivably, the matrix elements can be calculated in Excel, and the necessary matrix operations performed with an add-in such as Matrix.xla.

Some readers may expect references to multiple deprivation indices, some of which are based on ranked indicators. We discuss principles and differences vis-à-vis rapid needs assessment in a sidebar in the main text.

Severity and priority in Syria's J-RANS II

Assessments in Syria

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

the second Joint Rapid Assessment of Northern Syria (J-RANS II), extending the area assessed to 106 subdistricts (AWG 2013b).

We use material from this third assessment database. An important facet to know about this assessment is that in most sub-districts enumerators spent enough time in order to interview several groups of key informants. They used the same type of questionnaire in each meeting and ultimately created their personal synthesis for the subdistrict in yet another copy. Only these synthesis copies were use in debriefings and data entry. The variablility among informant groups within subdistricts is not accessible in the database.

Data architecture

The J-RANS II database is in single-record format. There are 106 observations, one for each of the 106 assessed subdistricts, and reflected in 106 records in one spreadsheet table. Thus comparable variables, notably severity and priority scores, for each sector, are held in separate (and differently colored) sets of fields side by side, rather than being on top of each other. We note this because the architecture created for the Yemen assessment, as we shall see, was different.

As mentioned, in subdistricts with multiple key informant interviews, the database record was filled with the synthesis version that the enumerators created from their notes. The questionnaires filled out with particular groups of informants, if they survived in hardcopy, did not result in multiple records being created on each assessed subdistrict.

Severity

How it was measured

The J-RANS II measured the severity of unmet needs in these formats:

Independently in each of the following sectors: health, food security, nutrition, shelter and water supply, key informants in sub-districts were asked to express their beliefs on a five-level severity scale. They were asked which of the following statements best described "the general status" of the sector:

Table 2: The severity scale used in the J-RANS II

1. No concern – situation under control
2. Situation of concern that requires monitoring
3. Many people are suffering because of insufficient [supply of goods or services]
4. Many people will die because [supply of goods or services] are insufficient
5. Many people are known to be dying due to insufficient [supply of goods or services]

The priority question was formulated as the sectors with "the most serious problems". Key informants were asked to prioritize five out of seven sectors in this list:

Table 3: The priority question in the J-RANS II

After these specific questions, we want to recapitulate. In terms of which sector poses the most serious problems, can you say which is the most serious, second most, third most, fourth most, and fifth most serious? I read you a list of 7 sectors:
[Priority Level. Rank 5: 1=first rank, 2=second rank, 3=third rank., 4=fourth rank; 5= fifth rank]

Health
Food Security
Nutrition
Water, Sanitation, Hygiene
Places to live and Non-Food Items
Education
Protection

To repeat: As elsewhere in the ACAPS terminology, the first question is understood as a *severity* measurement, the second as a *priority* measurement. The first aims at unmet needs in absolute (not looking at those in the other sectors), the second in comparative terms. The first produces ordinal, the second, under the Borda count interpretation, interval-level measures.

The lack of differentiation, a frequent problem with rating scales, as noted in the introduction, undermined the severity measurements here as well. They heavily congregated in the middle category, *"3. Many people are suffering"*. The percentages taken up by this category were as follows in the five rated sectors:

Table 4: Frequency of level 3, "Many people are suffering", in five sectors

Health	77%
Food security	76%
Nutrition	86%
Shelter and NFI	78%
Water supply	82%

The upper extreme, *"5. People are dying now"*, was used by respondents only in three instances, all with regards to health.

Over the entire sample of 106 subdistricts, the medians of the severity measures in all five sectors invariably were 3, or "Many people are suffering". This holds regardless of whether they were unweighted or population-weighted. The measure obviously does not discriminate at the aggregate level. Broken down to governorates, there is some rare variation. Thus, the median severity ratings for one sector in Hama as well as for four sectors in Lattakia were 2.

Despite the equality of the median severity scores across sectors, the assessment team made some inter-sector comparisons. These concerned the four sectors food security, health, shelter and NFI and WASH (nutrition was omitted). They were done in two ways. One table, in heat map fashion, visualizes the percentages of the accessed governorate populations for which the (subdistrict-based) ratings were 3 and above (from "many are suffering" through "many are dying now"). These are called the "affected populations".

Another table gives governorate-wise population totals separately for subdistricts rated "Many are suffering" (called the "at-risk" population) and for those rated "Many will die" and "Many are dying now" combined ("at acute risk") . The overwhelming impression that the table makes on the rapid reader is that the population *at acute risk from health problems* is much larger and more widely spread than those threatened by the problems in the other three sectors.

While it is legitimate and instructive to separate the at-acute-risk population statistically, the conclusion, based on this one measure, that *"far more people have acute needs in health than in any other sector"* (ibid.), is questionable. When we compare those figures to ones obtained by combining the two top sector priorities, the population for whom health is a high priority is still by far the largest over their analogues in other sectors (we also calculated the value for nutrition). But the proportions change. Notably, the 9 to 1 lead of health over food security in the severity measure dwindles to less than 2 to 1 in this priority statistic. This one is also based on signals from a much higher number of subdistricts.

Table 5: Subdistricts and populations in high severity or high priority, by sector

Sector	Severity: At acute risk		Priority: Sector with most and second most serious problems	
	Subdistricts	Population	Subdistricts	Population
Health	13	2,004,500	61	8,716,999
WASH	3	242,000	23	1,932,668
Food security	3	220,000	50	4,935,620
Nutrition	2	83,500	48	6,636,680
Shelter and NFI	1	55,000	7	584,919

The methodological conclusion is that the severity measure, as used in the J-RANS II, *may* have good validity - it measured what it was supposed to measure: unmet needs. It *may* also have good reliability - other enumerators and key informants would have returned similar values. But it discriminated poorly, particularly between the broad category "Many are suffering" and the rarely used upper categories "Many will die" and "Many are dying".

The question, therefore, arises: How should the severity scale be reformulated?

Can the scale be reformulated?

The version used

We first consider the form of the scale as it has been used, and then the nature of possible alternatives. We take the example of the one used in health:

Table 6: Health sector severity scale, as used in the J-RANS II

E4. Overall, which of the following statements describes best the general status of health in this sub-district? (circle 1 answer)

0. DNK
1. No concern – situation under control
2. Situation of concern that requires monitoring
3. Many people are suffering because of insufficient health services
4. Many people will die because health services are insufficient
5. Many people are known to be dying due to insufficient health services

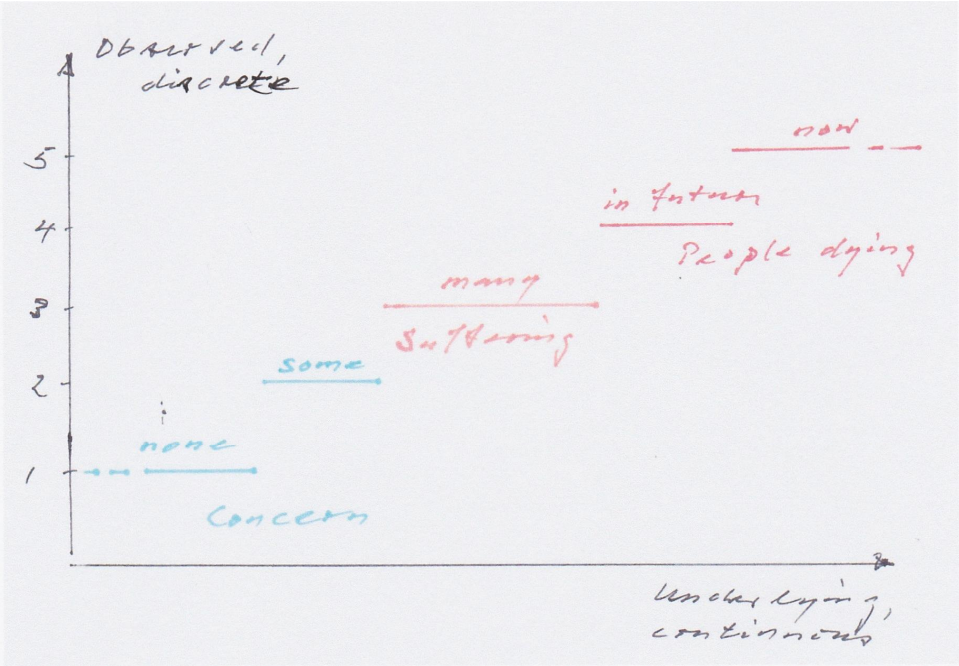
The options are based on three concepts: *concern, suffering*, and *death*. They are illustrated with auxiliary (control, monitoring) and causal ("because of insufficient services") terms. It would take Arabic speakers to evaluate the linguistic subtleties. Yet, it is obvious that the concepts differ in vagueness and emotional intensity. In a humanitarian disaster, the default is for every need area to be of concern, at least to the point of suggesting periodic monitoring. With almost similar universality, informants who have to speak for entire subdistricts will be bound to assume that "many people" are suffering because services are no longer sufficient.

The death of people is a much more concrete and verifiable event, or series of events. Deaths are countable, and even when this is not practical, observers will entertain numeric notions of those that already occurred. The problem may be more in the attribution to causes; in an armed conflict, the insufficiency of particular resources and services obviously is only one, and not always the principal, contributor to excess mortality. Incidentally, note another curiosity in the current severity scale: the two levels mentioning deaths are not differentiated by quantitative operators such as few / many / most, but by the temporal expectation: "will die" [in some not too distant future] vs. "are dying" [now].

One may speculate where respondents "anchored" (= set the comparison point for their understanding of) the scale. It is not very plausible that the pre-conflict situation (represented by the "no concern" option) should be the anchor. Rather, they may have determined their response in relation to the two "people .. dying" options. These are closer to what they have been going through lately. If their belief is that not many people will be dying, let alone are dying now, due to sector-related shortages, then they may cautiously settle for the umbrella category "many are suffering". Again, this is pure speculation until ACAPS debriefs some enumerators reporting on key informant conversations that dissected the meanings of such scale terms.

This figure summarizes the current architecture of the scale.

Figure 1: Latent unmet needs and semantics of the observed needs



Possible adaptations

Alternatives to the current scale range from minimal repairs to substantively different instruments replacing it. Whatever alternatives one contemplates, the current scale has the great advantage of having been tried out, supplying at least the kind of information that the assessment team found viable to present in the report. The alternatives are untested and therefore risky. In this section, we look at some of the minor adaptations that seem feasible. Other instruments that potentially could replace the current tool will be reviewed in a later section.

Minor change: Rewording the middle category

Thus, in a timidly small change, one may content himself by replacing the "many are suffering" with a new formulation that invites more restrained use. One could offer the option instead: *"People face shortages, but these are not life-threatening."*

The only thing that this achieves, however, is a clearer demarcation vis-à-vis the two death options. The wording clarifies that nobody is dying from causes directly linked to the sector. The distinction towards the lower levels is not clear. "Concern requiring monitoring" would not be justified if shortages were not experienced by some people.

More extensive changes

A more extensive adaptation would preserve the "one-question, several options" scale type, but would make changes in several or all of these elements: number of levels, key

semantic concepts, modifiers (quantitative, temporal, modal, etc.). A possible variant could work with more numerous distinctions within both non-lethal and lethal shortages:

Table 7: Severity scale - Possible reformulation

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

1. Most people are able to meet their needs
2. Some people are facing shortages, but these are not life-threatening
3. Many people are facing shortages, but these are not life-threatening
4. As a result of shortages, we will soon see some people die
5. As a result of shortages, some people have already died
6. As a result of shortages, many people have already died.

This six-level scale requires the respondent to assimilate three key concepts:

- The ability to meet needs
- Shortages that are not life-threatening
- People dying

Adjacent categories are distinguished by implicit or explicit moderator changes, except between 3 and 4, which introduces a more radical semantic change:

Meet needs: 1. Most do, 2. some don't, 3. many don't
Death: 4. Some [future] 5. some [present perfect], 6. many [present perfect].

Note also that this scale does not have a "neutral" middle scale, but requires respondents to make a "life or death" decision.

One could try to improve on that slightly, by re-arranging the key concepts and offering a quasi-neutral middle category:

Table 8: Severity scale - reformulated with 7 levels

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

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

The downsides are:

- Option #1 is implausible in a humanitarian disaster and only serves logical completeness;
- the respondents still need to process three key concepts (shortages, not life-threatening; death);
- as experienced with the current scale, the middle option is likely to be overused.

But, it is clearly demarcated against the lower and upper options. And, although it is untested, there are no obvious reasons why it should not function at least as well as the current scale did.

Priority

How it was measured

The question that was used to measure sector priorities was presented further above on page 12, in the segment on severity, in order to prepare the ground for the comparison of at-acute-risk populations and populations affected by priority sectors in Table 5 on page 13.

The priority score is a property of the sectors only. While it is possible to devise statistics based on the sector-wise severity scores that are meaningful to localities, it is not possible to use the priority score to differentiate among localities.

At data entry the ranks were recorded as they were, the first as 1, etc., with blanks left in the sector priority field when the sector was not ranked in the locality. In this form, the priority ranks are not optimal for analysis. They need to be reversed, with the first rank given a score of 5 (in this situation of five options) and the fifth rank a score of 1, and those sectors not chosen as a priority all of them zero. This way, they form an ordinal scale.

Analysis

The ordinality is a limitation. In the case of the priority score, it can be overcome under the assumption that the respondents understood the format. They were aware of their limited choices and of the ranking among them. The preference intensity differentials are not known in the individual case (i.e. for the key informants of a given subdistrict). The difference between first and second priority may not be the same as the one between second and third, etc. However, in the aggregate they should even out, perhaps with the exception of the unranked options, which were all dumped to the zero level of the scale.

This allows us to treat this data in analogy to an election system. The simplest system used in this context, and known from previous notes in the ACAPS toolbox, is the Borda count. Its premise is a ranking (originally of political candidates) of N objects in which the first preference is accorded N votes, the second N-1, etc. In many situations not all objects are ranked. As we have seen, in the J-RANS II ranks were available for five out of the seven sectors. The rest of the sectors remain unranked; all that can be said about them is that they are not among the first five priorities.

The ranks are ordinal. Statistically, the mean of the rank score has no interpretation. However, the Borda scores (as inverse ranks) may be summed in the understanding that communities "vote" for multiple candidates to be their priority sectors. These votes are weighted in the above manner. The "Borda counts" (Borda 1781; Benini 2011b; Wikipedia 2011a) is considered, if not an optimal, then a satisfactory ranking system, particularly on account of its simplicity. The "mean Borda count" above is the sum of "votes" for a sector divided by the number of sites. A system that does not rank all candidates is sometimes known as "modified Borda count". The major gain from analyzing priority data as Borda counts is the interval-level measurement.

Table 9: Priority scores - with and without population weighting - J-RANS II

Sector	Mean Borda count	
	Unweighted	Population-weighted
Health	3.62	3.74
Food	3.14	2.87
Nutrition	2.80	2.94
WASH	2.08	1.77
Protection	1.38	1.98
Shelter-NFI	1.35	1.14
Education	0.62	0.56

The results are robust to population-weighting at the top and bottom, with minor reversals in the middle of the rank table.

[Sidebar:] Priority ranking vs. ranking in multiple deprivation models

At this point, it is appropriate to point to a possible misunderstanding, particularly among readers familiar with certain types of multiple deprivation indices. These have become more common in the academic and policy literature on social exclusion, cumulative disadvantage, chronic poverty and the like. The methodological discussion is extensive and overlaps heavily with the social indicators literature.

Some of these models, for the normalization of their indicators, rank them. In subsequent steps, they weight and aggregate the ranked indicators, using appropriate transformations.

The point that we wish to emphasize here concerns the directions of ranking and of the subsequent aggregation, and how they differ from the handling of sector priorities in rapid needs assessments.

The main objective of multiple deprivation studies is to assign a measure of deprivation to each of the observed geographical or social units. Indices formed for this purpose mostly incorporate indicators of hard data, at ratio level. The reasons for reducing them to ordinal-level rankings are technical, primarily to make the indicators comparable and independent of their initial distributions

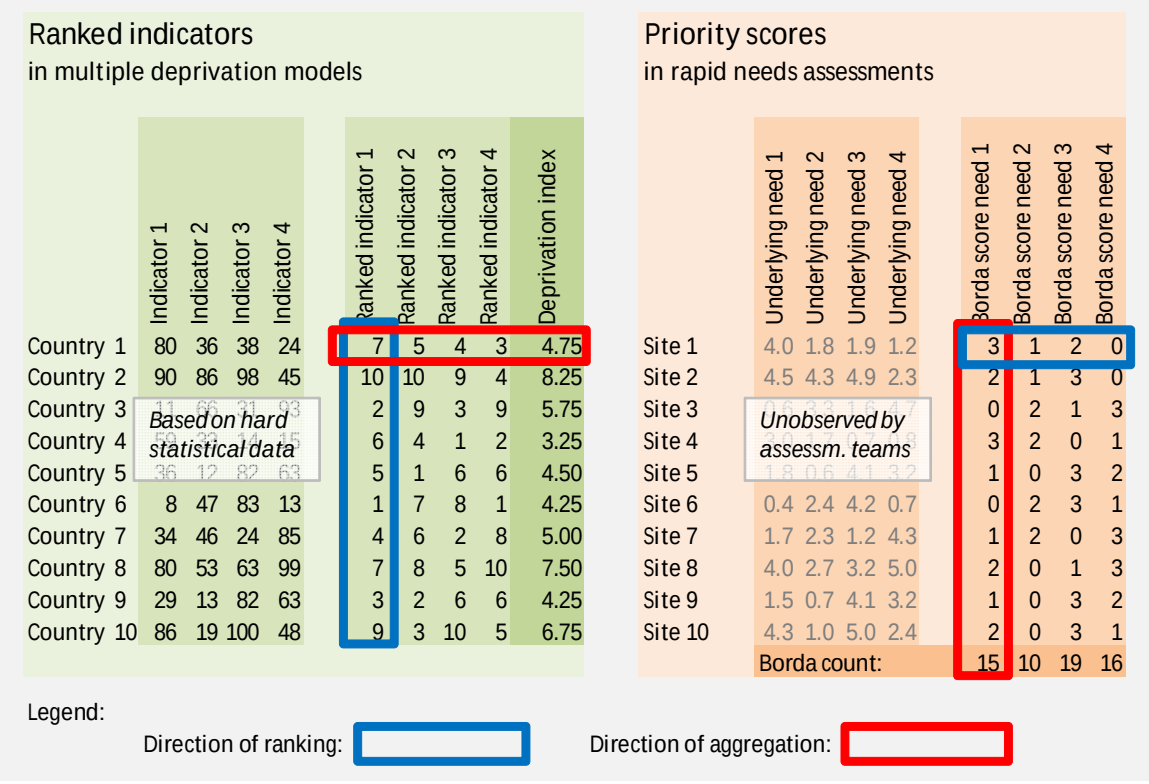
(particularly when there are no strong theoretical or policy reasons to retain the distributional information).

The ranking therefore proceeds within each indicator, over all the observed units. The aggregation takes place within each unit, over all the ranked indicators.

The approach to priority scoring in rapid needs assessment is the opposite. Key informants in a given unit (a site, district, etc.) rank sectors (each of which stands for a type of need). The underlying needs are not directly observed by assessment teams. The ranking, as already stated, takes place within each unit, over all the listed sectors (this is so even when the informants are allowed only a small number of options to rank). The outcome of interest is the aggregate priorities of the sectors; the sum of priority ranks for the individual sites is always identical (except for technical flaws) and thus is meaningless.

This diagram, using random scores for a fictitious set of ten countries / site and four indicators each, highlights the differences between the two approaches. The priority scores in this example use the Borda count interpretation (see above).

Figure 2: Ranking and rank aggregation, multiple deprivation models vs. rapid assessments



The Government of Scotland - and reportedly those of Wales and Northern Ireland too - have been publishing maps of multiple deprivation in small local units for a number of years. Methodologically, they were advised by the Social Disadvantage Research Center at the

University of Oxford, United Kingdom, whose 2003 report provides very readable step-by-step rationales (Noble, Smith et al. 2003)[4].

Severity and priority together

Observed aggregate values and correlations with priority scores

The J-RANS II severity scores have a median of 3 in all five sectors where severity was measured in this way. They remain the same when population-weighted. A sector ranking on this basis is not possible. Instead, the team differentiated sectors by the size of populations living in subdistricts with severity levels 4 and 5. This is the often referred-to Figure 14 on page 23, of which we reproduce a segment here:

Table 10: Populations at risk and at acute risk, by governorate (segment)

	Food security		Health	
Governorate	At risk	At acute risk	At risk	At acute risk
Aleppo	1,722,650		2,511,230	509,500
Al-Hassakeh	1,060,900		491,900	709,000
Ar-Raqqa	700,200	200,000	707,200	193,000
Deir-ez-Zor	1,916,919		1,605,519	331,400
Hama	1,380,150		1,175,950	141,600
Idleb	1,897,919	20,000	1,856,319	120,000
Lattakia	7,500		14,550	
Grand Total	8,686,238	220,000	8,362,668	2,004,500

The correlations between the severity and the Borda-coded priority scores for the five sectors are shown in this table[5].

Table 11: Correlations between severity and priority scores, in five sectors

Sector	Coeff.
Health	0.20
Food	-0.05
Nutrition	-0.06
Shelter	-0.35
WASH	0.81

[4] http://scotland.gov.uk/Resource/Doc/47032/0025597.pdf. The site for current reports is http://www.scotland.gov.uk/Topics/Statistics/SIMD.

[5] The statistic is the polychoric correlation coefficient (Kolenikov and Angeles 2004), with the subdistrict population as the analytic weight.

Only one sector returned a strong positive correlation - water and sanitation. In the sector that both measures identified as the one with the greatest needs in the aggregate - health -, the correlation at the local level is positive, but weakly so. We illustrate this with the population for the cross-tabulated scores in two sectors - WASH and shelter. First *WASH*:

Table 12: Population, by severity and priority levels - WASH sectors

Severity	Borda-coded priority (5 = first priority)						Total
	0	1	2	3	4	5	
1. No concern	435,500	3,000	0	0	0	0	438,500
2. Situat. of concern	2,651,800	0	422,000	7,050	0	0	3,080,850
3. Many are suffering	1,321,251	2,018,850	2,121,230	1,859,319	670,800	1,094,868	9,086,318
4. Many will die soon	0	20,000	0	55,000	0	167,000	242,000
Total	4,408,551	2,041,850	2,543,230	1,921,369	670,800	1,261,868	12,847,668

In the case of the negative correlation in the *shelter* sector, the cells marked middle-gray almost certainly represent measurement errors, and so likely do a number of observations with severity level 3, and Borda score 0:

Table 13: Population, by severity and priority levels - Shelter

Severity	Borda-coded priority (5 = first priority)						Total
	0	1	2	3	4	5	
1. No concern	0	0	3000	0	0	10,500	13,500
2. Situat. of concern	369,080	230000	1,510,550	0	0	96,000	2,205,630
3. Many are suffering	5,114,400	2,665,268	2,152,500	795,151	337,500	140,919	11,205,738
4. Many will die soon	55000	0	0	0	0	0	55,000
Total	5,538,480	2,895,268	3,666,050	795,151	337,500	247,419	13,479,868

[Note: The population totals for these two tables differ because the WASH and shelter severity scores each have one missing value, but in different records.]

Agreement at the aggregate level

As already shown in the table on page 13, by comparing sectors on populations

- at acute risk, by their severity scores, vs.
- with expressed first and second priorities

we find that

- both measures agree that the health sector creates the highest unmet needs
- both measures agree that the shelter and NFI sector faces the lowest unmet needs
- they disagree about the importance of unmet needs among the other sectors.

It bears repeating that the ratio between the sector with the highest population total and that with the second-highest differs enormously between the two measures, i.e. for severity: Health : WASH = approx. 2 million : 250,000 = 8 : 1; for priority: Health to nutrition = approx. 9 million : 7 million = 1.3 : 1. There is a trade-off between the overuse of categories in an absolute measure (#3 in severity) and the (by definition) perfect differentiation, but impossibility of independent interpretation in a purely relative measure (priority).

Within-sector "problems"

In the questionnaire, the elicitation of the severity score in each of the rated sectors is preceded by listings of "problems" (these are the equivalent of what the Yemen assessment called "key issues" - see later). In the health sector, for example, enumerators presented a list of 13 pre-defined health problems and, separately, one of nine health care issues (each with an additional text field for "Other"). Against each list, key informants selected up to five problems, respectively care issues. In addition, up to three priority health interventions were elicited. These were post-coded into approximately twenty types[6].

Formally, the severity scores do not depend on the patterns of response to these preceding questions. There is no evidence that the enumerators set the severity score in reaction to any such patterns (which, if true, constitutes one of the marked differences vis-à-vis the Yemen assessment format). It is plausible, however, that diligent enumerators, while reviewing the response in the questionnaires filled out in several meetings, made a personal "holistic" synthesis that helped them to choose the score if the scores offered by several key informant groups differed.

The long lists and the rules to check only a few options create statistical dependency among the resulting variables. Whether key informants interviewed in groups could make well-reasoned choices, is doubtful. And even if they did, the format may in part have obscured them. Thus, lack of medicine comes out as the top health problem. However, the shortage of functioning health care facilities is in the third rank; this may be due to the fact that it was the subject of multiple questions phrased in terms of availability and access. This diluted the response, so much so that it would not be correct to say, *on that basis alone*, that lack of medicine was a more serious problem than the shortage of facilities[7]. This format too should be revised in ways that break down the complexity and make for more clearly distinct pragmatic options.

One possible way to do this is to present a short list of distinct problems, identified by experts in the design phase, letting key informants select as many as they please, and then following up with an emphatic question: *"Our list may be incomplete. From your experience, are there any other urgent problems in this sector that you wish to bring to*

[6] "Post-coding" is an assumption. Persistent spelling variants rather point to initial field notes.

[7] Although, when taking into account also the priority interventions, it is confirmed that the lack of medicine dominates the list of problems.

our attention?" These additional nominations will have to be post-coded, to be identified as duplicates of some of the first choices or as genuine additions or elaborations.

Severity and priority in Yemen

The assessment in Yemen

ACAPS coordinated a "Joint Rapid Assessment of the Northern Governorates of Yemen" on behalf of CARE International in summer 2011. Several NGOs with an established presence in the country collected the data; the report was published in October (ACAPS 2011). The number of sites assessed was 43; however, one site record bundled several smaller sites; and three sites each appear in two records, one based on a key informant interview with a male group, and the other with a female group. Each site was populated with only one type of target group: the 43 sites include 12 host communities, 6 with returning IDPs, 21 with vulnerable IDPs, and 4 with other types of affected persons.

A key facet distinguishing this assessment is that *both community groups and assessment teams contributed severity ratings.* The community groups, under each of eight sectors (education, food security, health, livelihood, protection, return, shelter and WASH), would raise "key issues" in any number, name and order convenient and assign each of them a severity score (see below)[8]. They would suggest an intervention for each key issue. Once done with the listing and rating of all key issues across the sectors, they would designate first, second and third priority sectors, with a proposed intervention to go with each of them.

The assessment teams would categorize the key issues and the recommended interventions that the community groups had suggested. The categorized issues and interventions, together with the original severity scores, were preserved in the data entry. In addition, for each site and sector, the teams would offer a combined severity rating over all the enumerated key issues. This was known locally as a "synthesis" rating.

Altogether, the community groups had the teams note 1,033 key issues, which were recorded in 59 issue categories. They were mirrored in 63 recommendation types. Jointly, we find 146 distinct key issue-recommended action pairs.

As a result, both the communities' own and the assessment teams' severity perceptions are documented, the first as attributes of issues, the second as summaries by sector. This arrangement is different from the way the data were handled in Syria's J-RANS II, where only the enumerators' syntheses made it to the database.

[8] In its brief methodology section, the assessment report (page 11) makes a distinction between community groups (with whom qualitative interviews were held) and key informants (who sat in "structured, quantitative interviews"). The questionnaires, however, give the impression that it was mainly the response of community groups that populated the database. We are not going any deeper into this, assuming simply that community groups in Yemen fulfilled similar functions as the key informants did in Syria.

Data architecture

The coexistence of two kinds of severity ratings with the sector-level priorities and with site-level background information produced a more complex data architecture, with multiple records per site. The initial spreadsheet setup was largely motivated by data entry convenience and was not optimal for the analysis. ACAPS later published a reformatted version, together with data management and analysis notes (Benini 2011a, 2011b).

Severity and priority

Describing the treatment of severity and priority in the Yemen assessment is easier when we give a look at a screenshot of the original data capture interface. This figure shows part of the data entered for site #1.

Figure 3: Yemen assessment data arrangement (segment)

1		Sector	Problems Identified in priority order (max 5)	Severety of Needs Rank				Total	Recommendations for Intervention
Governorate	Amran	Livelihood	Income/employment			x		3	Income/Employment
District	Amran		Cash for basic services			x		3	Cash programming
Site Name	Alhboba-BeirHirab		Lack of skills		x			2	Vocational training
Site category	Village/part of town							NA	
Urban Rural	Urban							NA	
Total # populatio	11680								
Total # IDPs	3240								
		Synthesis				x		3	
		WASH	Water supply/management			x		3	Provision of water, tanks
Target Group	Vulnerable IDPs		Water sources		x			2	Rehabilitation of sources
Gender	Male		WASH NFIs status	x				1	Provision of Hygiene Items
Average age	30		Water quality			x		3	Water treatment (filters)
								NA	
Priority 1								NA	
Sector	Cross-cutting								
Intervention	Provision of support to nev	Synthesis				x		3	
AT comment		Shelter	Shelter Security/Condition			x		3	Shelter repairs
			NFI status					NA	Provision of NFIs
Priority 2			Heating/Cooking		x			2	Provision of heating fuel
Sector	Health							NA	
Intervention	Access to life saving emerg							NA	
AT comment									
Priority 3		Synthesis				x		3	
Sector	Food Security	Food Security	Food availability			x		3	Food supply
Intervention	Food rations and gas for th		Food accessibility			x		3	Registration, advocacy
AT comment	Winterization (blankets, et							NA	
								NA	
								NA	
		Synthesis				x		3	

The segment on the left side holds administrative, demographic as well as priority sector and intervention information. However, what dominates the image, and in fact holds the key variables in the assessment, is the color-coded severity scale to the right. Yemen employed a four-level scale, made graphic by the color ramp from green to red. What matters here is the interpretation of the scale levels, as in this table[9]:

[9] Assessment report (op.cit., page 103), and "Questionnaire B: (version 1.5): IERP 2011 - COMMUNITY GROUP DISCUSSION RECORD SHEET" (page 11), which the participating NGOs used in the field.

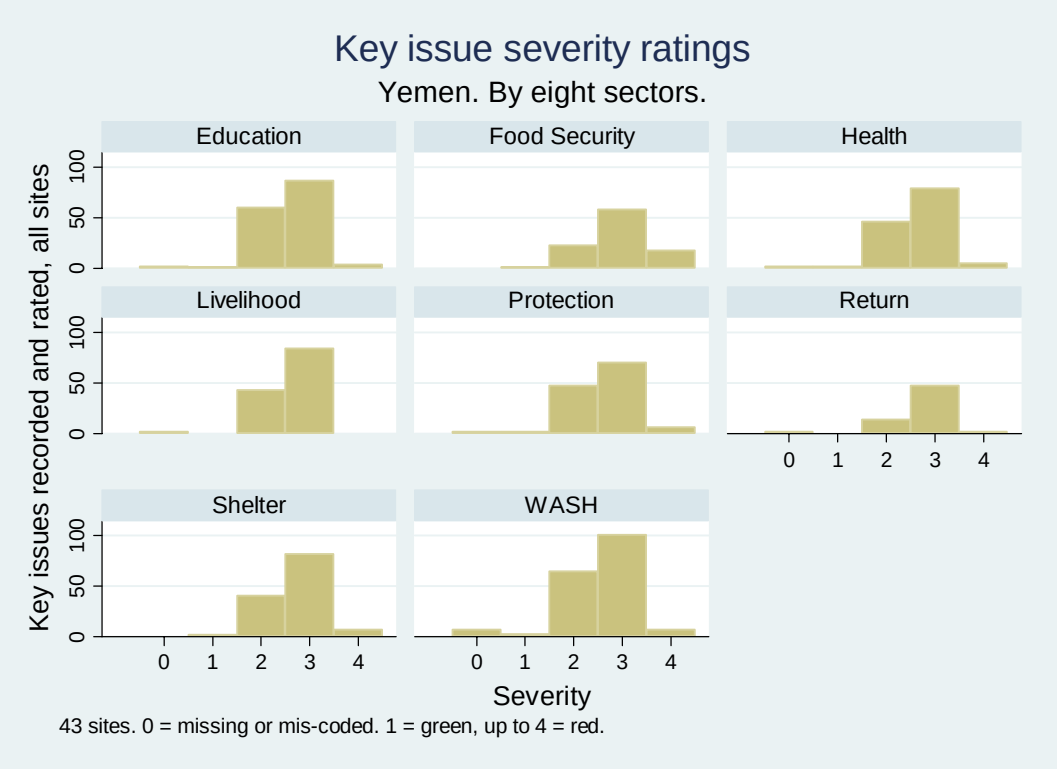
Table 14: Severity scale - Yemen assessment

Low	Relatively normal situation (or good data) or local population able to cope with crisis; no further action required
Medium low	Situation of concern, lack of data/unreliable data: further assessment and/or surveillance required
Medium high	Situation of concern, serious risk and lack of data/unreliable data: further assessment and/or surveillance required
High	Severe Situation: Immediate intervention required *to save lives*

For the analysis, the levels were coded from 1 (for green) to 4 (for red).

With four levels, the scale is user-friendly, but does little to aid differentiated judgment. Also, the interpretation mixes two dimensions, severity (expressing conditions that threaten life) and uncertainty (felt as lacking or unreliable data).

Figure 4: Key issue severity ratings, Yemen, by sector



In the event, the severity ratings of issues (by the community groups) as well as those of sectors (by the assessment teams) congregated significantly at level 3 (orange: Situation of concern, serious risk). Community groups assigned this level in 60 percent of their issue evaluations; for the sector evaluations by the teams, the proportion was similar (63 percent). Level 1 (green: relatively normal situation) was rarely used by the community groups, and never by the teams - presumably because such situations remain below the

recognition threshold in the elicitation of issues. As a result, the median severity score for all sectors uniformly was 3 in issue ratings by community groups as well as in sector ratings by teams. The same uniformity prevailed when median severity scores were computed for governorates.

The assessment report emphasized sector-wise and target group-wise analyses of issues and recommendations, with no attempt to compare sites or sectors by severity levels. It is not obvious how the "severity rankings" ("ratings" would have been appropriate) were calculated for the numerous graphs of priority needs by target group and/or governorate. Anyway, the distillation of statistically correct measures suitable for comparisons across sectors or sites would have been challenging, given the low differentiation of ratings.

Sector priorities produced better differentiation. The mean Borda counts for two of the three sectors with the highest needs, and for the three with the lowest, are fairly robust to population-weighting[10].

Table 15: Priority scores - with and without population weighting - Yemen

Sector	Mean Borda score	
	Unweighted	Population-weighted
Food security	1.67	1.39
WASH	1.44	1.04
Livelihoods	1.00	1.19
Health	0.81	0.42
Shelter	0.67	1.12
Education	0.19	0.11
Protection	0.00	0.00
Return	0.00	0.00

Analytic pros and cons

The agreement between community groups and assessment teams was low. This, at least, is the impression when we compare, for each site-sector combination, the "synthesis" severity ratings by the teams to priority scores implicit in the choice of three sectors that each community group designated as its priority sectors.

[10] The Borda count was mentioned earlier. In the Yemen assessment, community groups could rank only three sectors as priority sectors. The rest of the sectors remain unranked.

Table 16: Agreement between teams' severity ratings and communities' sector priorities

Severity in *assessment team* synthesis	Priority sector for the *community*		Total
	No	Yes	
2	58	19	77
3	105	59	164
4	12	19	31
Total	175	97	272

The agreement between community priorities and assessment team judgments is minor. Only in a fifth of all situations where key informants ranked a sector as one of three priorities did the assessment teams grant a "severe" grade (19 out of 97). And, only for two thirds of the "severe issues" (19 our of 31) had the key informants indicated that these sectors were among their priorities.

The low correlation between the judgments of the two sources - community groups and assessment teams - may disappoint some, undermining their trust in the validity of the measures and the reliability of the data. However, the Yemen format offers its own opportunities for meaningful severity/priority analysis.

The community groups nominated key issues in each of the eight sectors. These were narrated in free form, not chosen from lists. The assessment teams categorized the issues for data entry, utilizing 59 different issue formulations. These were further reduced in a secondary analysis later. By forming broader categories, the sector-key issue combinations could be reduced to 18. By filtering to key issue instances within the sectors that the concerned community groups considered priorities, the list of key issues was further restricted to 12.

At this point, rules were required to determine how to rate the 18 key issues. In the Yemen data re-analysis, three such rules were proposed:

1. first adopt a rule of **prudence**, to say that among the 12 problems from priority sectors any rated as "severe" (level 4) by more than one community should be highlighted red for attention. This discourages the inclusion of problems that, while from priority sectors, were not typically rated "severe".
2. next adopt a rule of **fairness** for the distinction between "serious risk" (level 3) and "more data needed" (level 2). Prior to filtering to priority sectors, mark any problem orange if it was not yet marked red and either had at least two severe instances or more than 20 instances of serious risk (20 being approx. half of the number of sites [43]).

3. The remainder of problems would be highlighted yellow. In other words, we assume that none of these reclassified problems warrants a "no major problem" score (green)[11].

For Yemen, the resulting sector-key issue table can be formatted like this:

Table 17: Summary key issue importance, Yemen

Sector	Issue	Importance	Sector	Issue	Importance
Education	Education, access		Protection	Conflict, crime and violence	
	Education, delivery			Safety and security	
Food Security	Food, effective access		Return	Financial	
	Food, quality			Safety and security	
Health	Health care capacity		Shelter	Financial	
	Health condition			Safety and security	
	Human resources		WASH	Hygiene	
Livelihood	Financial			Sanitation	
	Human resources			Water	
Legend: Red = high importance; orange = medium importance; yellow = low importance.					

Note: Condensed from 59 issue categories with 984 community group severity ratings from 43 sites.

Such rules are productive to signal priority problems. They can vary from assessment to assessment, and from analyst to analyst, as long as they are explicit and offer some rationale. In this case, they were invented and applied in order to make synergistic use of the sector priorities, key issue re-categorization as well as severity ratings. A similar approach might prove difficult with assessment formats that pre-categorize key issues (as was done in Syria). The number of issues might be too high, and the numbers of instances in most of them too low, to define filters for reasonably robust groupings by importance.

Comparing the way severity and priority were measured in the Northern Yemen and Syria J-RANS II assessments, three points stand out:

1. In both assessments, the severity scales did not differentiate enough to establish sector priorities.
2. The priority measure established clearer sector differences.
3. Largely because of 1., the correlations between severity and priority scores were lower than expected. In Yemen, subsequently sector priority and key issue severity scores were combined to produce a list of most important key issues.

We now return to more basic considerations that for the most part abstract from the Syria and Yemen situations, while sometimes referring back to them.

[11] The procedure is shown in greater detail in Benini (2011a: op.cit.).

Severity and priority - General considerations

Severity measurement

In general

As argued before, "severity" as a concept needs to be explicated in terms that take it closer to a measurement logic. There are various options. One is to tie it to observed events, such as deaths and displacement. This would work to the extent that future deaths and displacement can be predicted, or at least the risks validly estimated, on the strength of past events. Alternatively, we may associate a concept that from the start is more dispositional in nature. The idea that severity expresses the degree of "unmet needs" is worth pursuing. Unmet needs exist today and have consequences tomorrow, by prompting acts to mitigate them or, failing that, by causing further deterioration in affected groups.

There is a measurement rationale in basing severity on unmet needs. At a basic level one may argue that unmet needs are not directly observed, but are inferred indirectly through acts such as interview questions and the evaluation of the response. One may also assume that underlying needs can be expressed, for each sufficiently distinct domain, as continuous variables whereas the observed measures may be continuous (e.g. anthropometry) or, more practically in rapid assessments, discrete and ordinal.

This view may be disputed, for both the latent as well as the observed notions. Unmet needs in a sector may be difficult to reduce to one dimension. In the J-RANS II context, the health care needs of injured persons and those of people exposed to other risks may be highly different in nature, and measures aggregating them into one expression of unmet health sector needs in an area may be hard to defend. At the observation level, needs may be operationalized in continuous or count variables, such as through estimates of caseloads, that are stronger than ordinal measures.

While noting these difficulties, we cannot resolve them here. We continue the assumption that unmet needs can be thought of as one-dimensional continuous latent variables, and can be measured through ordered categorical constructs.

Alternatives to the J-RANS II and Yemen scales

Some of the challenges to the severity scales, in the J-RANS II and Yemen versions, were discussed further above, notably their failure to usefully differentiate, as seen in the frequencies of the response. The failure most likely originated from an insufficient number of levels (four in Yemen, five in the J-RANS II), from mixed semantics, and, in the Yemen case, from the fact that key issues were selected by community groups, which brought to recognition only those above a minimal intensity.

In this section, we consider alternatives that go beyond mere changes in semantics. Rather they concern the form of the tool. We distinguish primarily between measurements that employ one question (with explanations, lists, visual supports, as needed) and those that collect the response to several questions. These produce different

variables, which are statistically independent. They will then have to be combined in one quantity of interest per assessed unit and sector.

One-question tools

Verbal tools: Vignette-driven questions

In theory, one can elicit coarse estimate of the direction and amount of change in unmet needs by depicting a reference situation. Short verbal sketches using familiar topics and language provide a strong image. In this image, respondents can anchor their knowledge of their own current situation. Example:

Table 18: Text of a fictitious vignette-based question, Syria

Health sector:

In a previous assessment in northern Syria, two months ago, it was found that

- only one quarter of the population had access to functioning hospitals,
- half of the population in need of care had access to physicians,
- and only half of the pharmacies were still open.

Compared to that situation, how would you describe the health care situation in this subdistrict currently?

- It is a lot worse.
- It is somewhat worse.
- It is essentially the same.
- It is slightly better.
- It is a lot better.

[Additional question:] What makes you think so?

Vignette questions have strong pros and cons. If handled well, they can produce valid information on a complex topic. They are well-suited to produce response in several parts, some provoked by follow-up questions. In the example here, one works for a combination of a standardized ordinal response and a verbal, qualitative elaboration.

However, vignettes need to be pretested, particularly when translation is involved, and the handling of the qualitative part can be as distracting as it is instructive, both at the interviewing and the analysis stages.

Moreover, the consumers of the assessment may not be impressed with the kinds of statistics that follow from such a setup - e.g. *"in xx percent of all accessed subdistricts, the health care situation was felt to have deteriorated in comparison to what it used to be a typical situation at the time of a previous assessment, etc."*. In other words, the metric would be considered too weak.

The vignette approach here is relatively trivial compared to the factorial design of Rossi et al. (Wikipedia 2013d). We do not think it is feasible to vary vignette parameters during

one and the same assessment. The literature does not indicate that such methods have already been experimented within needs assessments after disasters, except perhaps in mental health evaluations and in other rare and atypical circumstances (Sayre 2006). We therefore mention vignettes for completeness, but do not recommend their use (yet)[12].

Numeric tools: Ladders

Ladders are scales presented to respondents as the analogue to a physical object - the ladder. The difference vis-à-vis a scale is that the rungs of a ladder do not all have an explicit verbal description. They may or may not all have a clear quantitative interpretation, and some are nothing but an ordinal scale in which the interpretation of the interior rungs is left to the respondent.

The most widespread applications are in human welfare and human wellbeing measurement. In this realm, the image of a person climbing up and down the ladder has an immediate and well-understood connection with social stratification and mobility. A version used in subjective welfare research is the "Economic Ladder Question (ELQ)" (Ravallion 2012: 7):

> "Imagine six steps, where on the bottom, the first step, stand the poorest people, and on the highest step, the sixth, stand the rich (show a picture of the steps). On which step are you today?"

Note that only the endpoints are interpreted. The respondents who feel they are neither rich nor very poor have to determine for themselves which of the middle steps suits their situation. An interesting question regarding ladders in general is whether, given enough steps, the interpretation must still be ordinal at best, or whether the respondents make use of the levels in a way that permits interval level analysis (though not necessarily ratio level, since a meaningful zero point may be unavailable). At least from one area, life satisfaction research, Cojocaru and Diagne (2013: 3) report that interval-level use is justified:

> Ferrer-i-Carbonell and Frijters (2005) examine the more stringent assumption of cardinality, i.e. that the difference between responses 2 and 3 on the satisfaction scale is the same, for instance, as the difference between 6 and 7. Relying on data from the German Socio-Economic Panel (GSOEP) they look at differences between ordinal and cardinal models of life satisfaction using the 11-step response to the following question: *"How happy are you at present with your life as a whole? Please answer by using the following scale in which 0 means totally unhappy, and 10 means totally happy."* They find that results are largely unaffected by the choice of cardinal vs. ordinal specification.

It is, of course, an open question whether this result holds also for measures of unmet needs that emulate a ladder format.

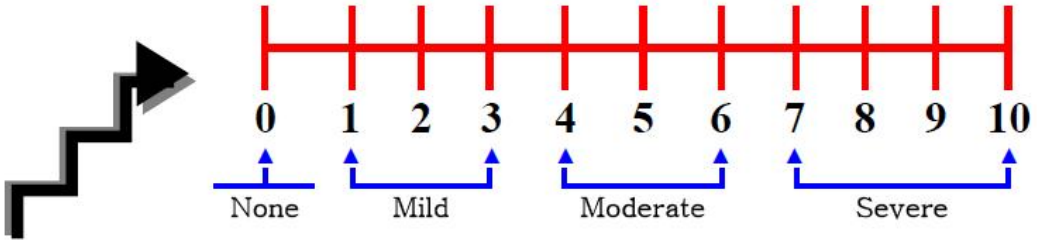
[12] King, Murray et al. (2004) make a powerful argument for the use of vignettes in intercultural survey situations. Their measurement-theoretical approach is worth studying also for rapid needs assessment.

The situation would be different if each of the ladder rungs had a clear quantitative location. For example, suppose that respondents were to imagine *"ten steps, one for each income decile in the population, with people in the poorest income on the bottom step, and those in the highest decile on the top.."*. Then, naturally, all steps in-between would have an analogous quantitative interpretation, e.g. the fifth step would hold the people in the fifth income decile, etc. This is hardly practical in rapid needs assessments, for lack of a dominant interval-level indicator and, even if it existed, for the difficulty for respondents to know in which decile their area falls.

The pain scale as a possible inspiration

For completeness, let us mention also certain types of pain scales. They hold a middle ground between ordinal scales with complete verbal descriptions and ladders with partial interpretations. This version:

Figure 5: National Health Institute pain scale



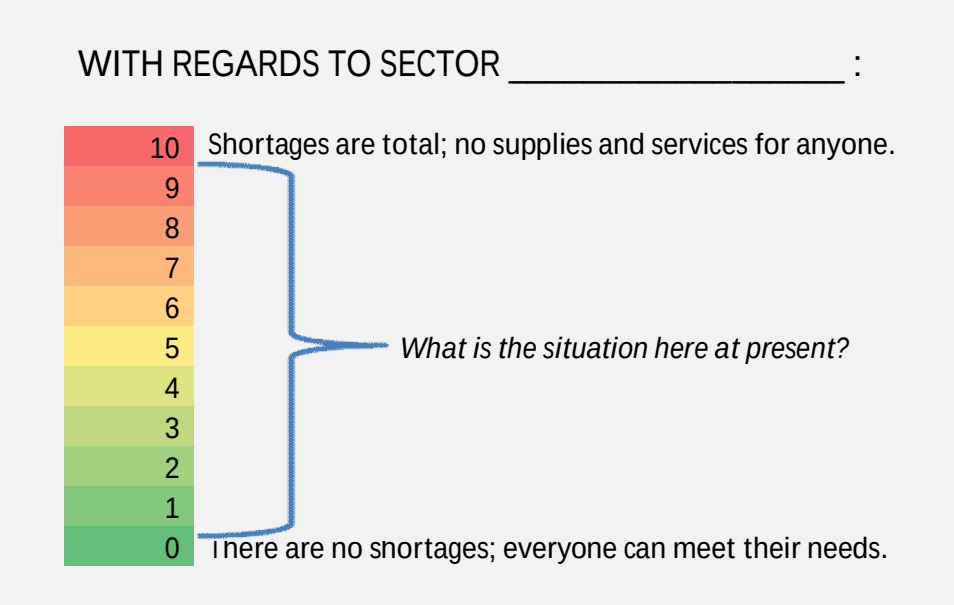
was taken from a NIH visual (NIH 2003-07). Importantly, nurses help patients to understand mild pain (it does not interfere significantly with activities of daily life) moderate pain (it interferes significantly) and severe pain (patients are unable to perform them). In other words, the interview protocol provides for additional interpretation.

One could consider an elaboration for the measurement of unmet needs. Particularly among assessment teams with high education levels and overlap with medical professions this might conceivably work well. We would need to find guidance for the understanding of "mild", "moderate", "severe" in ways that make measuring unmet needs possible across several sectors.

A ladder with endpoint descriptors only

This eleven-step ladder, with no descriptors except at the end points, may induce more variability across need domains.

Figure 6: Visual for an 11-rung severity ladder



It will, however, make the definitions of "at risk" and "at acute risk" areas more arbitrary.

The ladder, with colors and minimal descriptors, is proposed as a visual, possibly laminated and A4 size (or larger) for use in group discussions.

Multi-question tools: Item collections

Scales and ladders impose one-dimensional response. This constraint may obscure important distinctions in the nature of needs as respondents try to situate perceived severity in the set of options given them. Ideally, one would want to use instruments that permit the discovery of several dimensions, or at least to evaluate the quality of the one-dimensional measurement.

There are, in theory, alternatives that avoid one-question-based scales and ladders. We can mention three; there may be more. This para is brief; we do not think these tools viable because rapid assessments will not have the time for minimal pre-testing.

An exception could be made if an already tested and validated item pool from other disaster research were available. Even then the risks of inadequate translation and training might be too daunting.

In the Syria context one could make a case that ACAPS should expect to support more assessments in future, that the working group has accumulated considerable expertise and familiarity with context and tool, and that the assessment teams are highly educated and represent various technical fields. Some of them could therefore be asked to help devise sector-specific needs measurement instruments of one or the other kind below.

Dichotomous items

One could, for each sector, develop a small number of diagnostic yes/no questions that plausibly relate to different levels of service provision. In mildly affected areas, informants would answer most of them with "yes". In severely affected, service-depleted areas, "no"s will dominate. In quick and dirty analysis, the number of "no"'s would characterize the severity level in the sector in point. In more sophisticated analysis, and with enough areas returning complete answers, the severity contribution of each item could be estimated (as a so-called Rasch scale) (Wikipedia 2013c). The questions would need to be pragmatically distinct and sector-wise meaningful[13].

The following example scale is for the food sector. It has been adapted for key informant use from Coates et al. (2006: 1442S: Table IV), who designed the scale for rice-growing communities.

Table 19: Sample scale of dichotomous items

Sub-scale	Item
Inadequate quality	Do you feel most people in this area cannot afford to eat properly?
	Do most families cook the same food day after day?
	Do most of them eat wheat (or another grain) although they want to eat rice?
	Do many people eat rice starch because they lack money for food?
	Are most people not able to cook hot rice?
Insufficient quantity	For most people, does the food they buy not last, and they don't have money to buy more?
	Do most people eat less than you think they should because they don't have enough money for food?
	In many families, do some members eat less food so that there will be more for the rest of the family?
	Since the last harvest, did many families reduce the number of their daily meals?
Socially unacceptable	Do you ever observe women working in the fields with men?
	Do many people eat wheat gruel because there is no money for other food?

[13] E.g., regarding food: "Not enough access to food due to lack of money" vs."Price increase of basic food items" (J-RANS II report, op.cit., p. 36, Figure 44.) are semantically distinct, but not pragmatically so - incomes are not sufficient because prices are high, and prices are too high in view of low incomes.

Whether this scale consists entirely of pragmatically distinct items may be debatable, but the basic structure of combining yes/no questions should be obvious. The same authors report that the US Department of Agriculture developed a 18-item food insecurity scale that proved to conform to a Rasch model[14]. Similarly, the Humanitarian Emergency Settings Perceived Needs Scale (HESPER) might be scanned for items suitable to be adapted for sector-wise scales administered to key informants (WHO and King's College London 2011).

Likert scales

The Likert scale is the result of analyzing data collected in the form of Likert-type agree-disagree questions. Critically, the questions all have to refer to a *common object*, which in our case would be the unmet needs in the sector in point. One of the assumptions underpinning this model is that, although the individual question may produce an ordinal response, if response to enough items is generated, the sum of scores can be treated as an interval-level measure.

The challenge in the unmet-needs measurement context is the generation and handling of a sufficient number of meaningful items for each sector. These items would need to be universal enough to function across different units of the assessed disaster region and possibly also across points in time. Moreover, only a comparative analysis would be possible ("Area A is worse off than area B"), but absolute inference in the sense of "Area X now has Y level severity in Z sector" would not be feasible on the basis of agree/disagree-type response sets.

Formula-based index

The response to a small number of questions within a sector could be considered indicators, and these could be combined in an index, assigning each of them a specific weight. For example, the price of bread, the fact that flour deliveries are the first food security priority, and the lack of cooking fuel could inform a food severity index.

Problems with arbitrary weights, missing data, and too narrow scopes of indicators discourage this approach.

[14] "The notion of *orderliness* or *predictability* to the food insecurity response has influenced both U.S. and developing country attempts to measure food insecurity along a range of severity. In the United States, the observation of orderliness drove the USDA's choice of statistical model, the Rasch model, to guide the development of the U.S. 18-item food insecurity scale. During the Sahelian famines of the 1980s, monitoring changes in the progression of "coping strategies" in the face of acute shocks was thought to provide potentially useful information that could trigger a humanitarian response. During the late 1980s, the measurement of coping strategies began to be codified into food security and nutrition monitoring and early warning systems and vulnerability assessments" (p. 1439S-1440S) and the references given there.

Interaction of severity and priority measures

General considerations

The severity scales are administered and analyzed as independent measures for each sector. The priority scores are relative measures expressing beliefs that the unmet needs felt in sector A are more serious than those in sector B. The verbal stimuli used in the measurements may differ. For example, the J-RANS II questionnaire introduced the severity questions asking about the "general status of" welfare good X (health, "ability to eat", etc.); when prioritizing sectors, it wanted to know which sectors were posing the most "serious problems". We assume, however, that respondents related both questions to the same concept of unmet needs, differentiated by sectors. This is an assumption - no more! -, which we uphold as long as there is no evidence for significantly different concepts. If correct, then, severity and priority measures have the same objects, albeit different metrics.

The basic intuition is that if the underlying unmet needs in sector A are higher than those in sector B for the sample average, then, in the absence of measurement error, both the ordinal severity measure and the (ordinal or Borda count-interpreted) priority measure will reveal A > B for the sample. The correlation between $I[S(A_{i)} > S(B_i)]$ and $I[P(A_i) > (B_i)]$ - with i for individual units such as subdistricts, I for indicator variables, S for the severity score, P for the Borda-coded priority -, will not be perfect because the categorical severity measures lump together some units that have different underlying values. But it should be strongly positive.

What happens when we have more than two sectors? The answer is not straightforward. It is not intuitive whether the priority result A > B can be upset by the intervention of other sectors C, D, etc. priorities. The mechanism that may subvert the expected result is known as Arrow's Paradox (Wikipedia 2013a). The Borda count does not protect against these so-called "irrelevant alternatives". If the distribution of priority votes for C, D, etc. does upset the A > B relation for the sample, then we should also expect a weakening or even reversal of the correlation between $S(A_{i)}$ and $P(A_i)$, between $S(B_{i)}$ and $P(B_i)$, etc.

Measurement error further complicates the relationship between severity and priority. Generally, measurement error attenuates correlations towards zero; it does not reverse them unless the errors themselves are correlated. For these reasons - interferences by the other sectors' preferences and error - one should expect relatively low correlations between severity and priority scores. Still, if the differences in the means of the underlying needs variables are significant, the aggregate order of sectors will likely be similar on both measures.

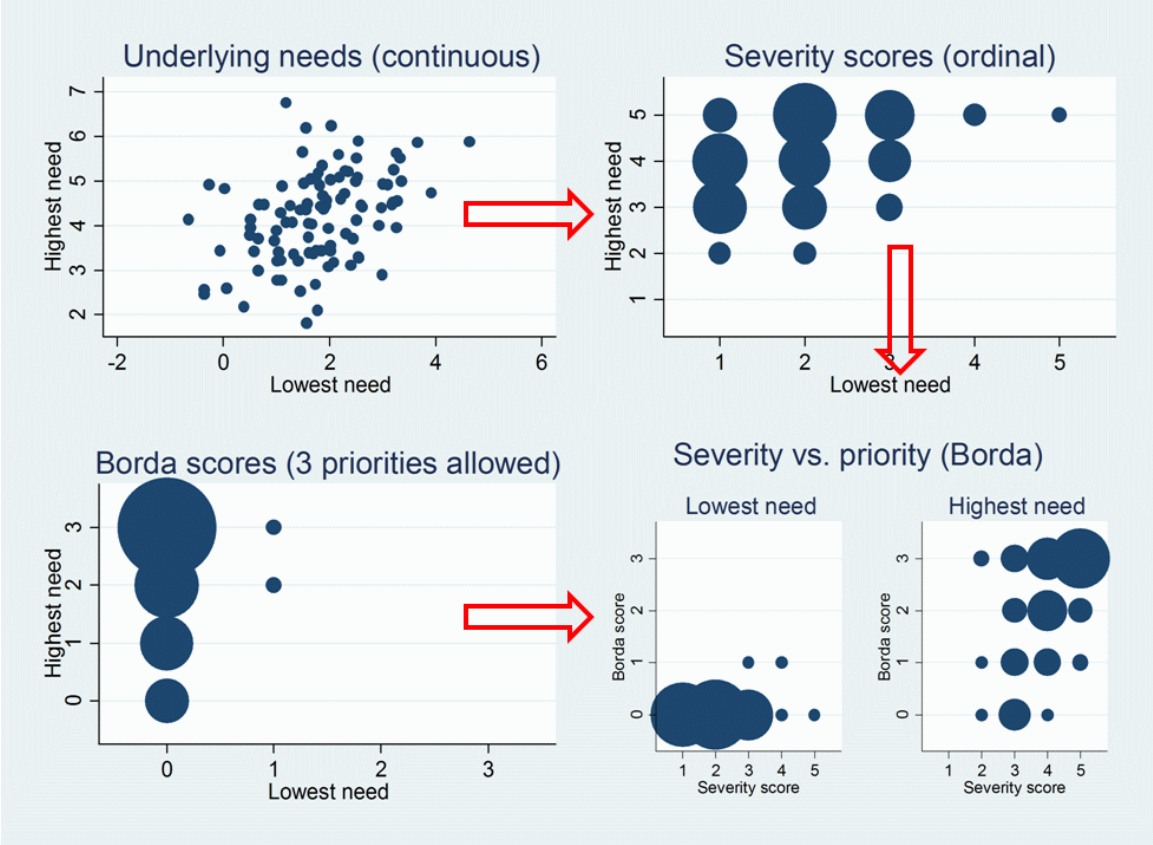
Simulation results

Because of the insufficient differentiation in the severity scores in those two assessments, we resort to artificial data. We explore the relationship between severity and priority with a simulation model which, we hope, readers will consider plausible.

We simulate underlying unmet needs in seven sectors for 100 communities, with means separated by equal distances across consecutive sectors, identical variances, and identical weak correlations (0.40). Severity scores on a scale from 1 to 5 are then computed by mapping the unmet needs scores, by intervals, to integer values. Borda type priority scores are formed on the order of the unmet need values, with the three highest needs scored 3, 2, and 1 and all the others 0. The simulation code is appended.

This four-panel chart illustrates the process from the simulated underlying unmet needs to the correlations of severity and priority scores. It exemplifies the connection for two needs - the lowest and the highest (the priority scores, of course, incorporate also the information on the other five needs).

Figure 7: The process from underlying unmet needs to severity and priority scores



Results when there is no measurement error

It is obvious from the lower right-side panel that we may expect to see a modestly strong correlation between severity and priority in the highest need, and a very low one in the lowest need. In fact, in the absence of measurement error, the rank order correlations between severity and priority are:

Table 20: Correlations between severity and priority scores (simulated data, 7 sectors)

Need	Corr.
Need #1	0.22
Need #2	0.34
Need #3	0.51
Need #4	0.45
Need #5	0.67
Need #6	0.51
Need #7	0.50

Even at the higher end (needs #6 and 7) the correlations are not very strong. This is so for the reasons mentioned earlier: truncation at the extremes and rounding of the severity score; multiple zero codings and dependence on the other needs variables in the priority score. Thus, even under ideal conditions (perfect reliability and high validity of both measures), we should not expect high correlations. For setups similar to seven sectors and three priority options, and similar variability of needs across sectors, rank-order correlations in the neighborhood of 0.50 are at the high end. In practical terms: if we draw bubble plots of severity vs. priority scores for a given sector, there will be sizeable bubbles also outside the "main diagonal". This is easy to spot in the two half-panels in the lower right corner of the chart: there are sizeable bubbles for the severity score - Borda count combinations (3, 0), (4, 1), (5, 2) below the diagonal as well as (3, 3) and (4, 3) above it, all of them penalizing the correlation.

With measurement error

Realistically, we must expect that both severity and priority are measured with significant error. This is so because key informants find it difficult to assign a discrete verbal or numeric option to the bundle of unmet needs that they feel within a given sector, and all the more so to use those options consistently across several sectors. In addition, there may be errors made by assessment workers, at the interview or later stages.

We introduce various levels of measurement error into our simulation model, ranging from no error to unrealistically high error levels. The question of interest is about the impact of errors. How do they affect the distributions of severity and priority scores? How is this perturbing findings at the aggregate (for the entire sample of assessed communities) and at the individual level (the correlations among scores over the assessed communities)? We ran simulations using different error levels, each with 100 replications (of which each was initialized with a different random seed).

Sample averages: *Without error*, the priority scale discriminates more keenly at the upper end - it assigns different sample-wide median scores to the highest need (3, meaning first priority, to #7) and to the second-highest (2 for second priority, to #6). On severity, both of these needs have median scores of 4. The opposite is true, by design, at the lower end, where the severity score discriminates better (at this end, most instances are such that they simply result in a priority score of 0).

But already with *realistic error levels*, the sample-wide medians of the priority scores can vary, with the result that needs #6 and 7, or #5 and 6 for that matter, can produce the same medians[15]. There is no variation in the medians of the severity scores in #6 and 7. Both remain at 4.

At higher levels of error, the medians also of the severity scores grow variable in some needs. Note that the highest need is virtually unaffected; it does not descend below level 4 (which means "many people will die" in the J-RANS II). However, things get more fluid between needs #5 and 6. The situation of the priority scores has not changed much further.

Correlations: Correlations of the rank order kind were computed, separately for each need, between severity and priority scores, over the 100 locations in the model. For illustration, we table the mean Spearman's correlation coefficient in need #7 over the 100 replications for all combinations of error levels.

Table 21: Correlations between severity and priority, by levels of measurement error

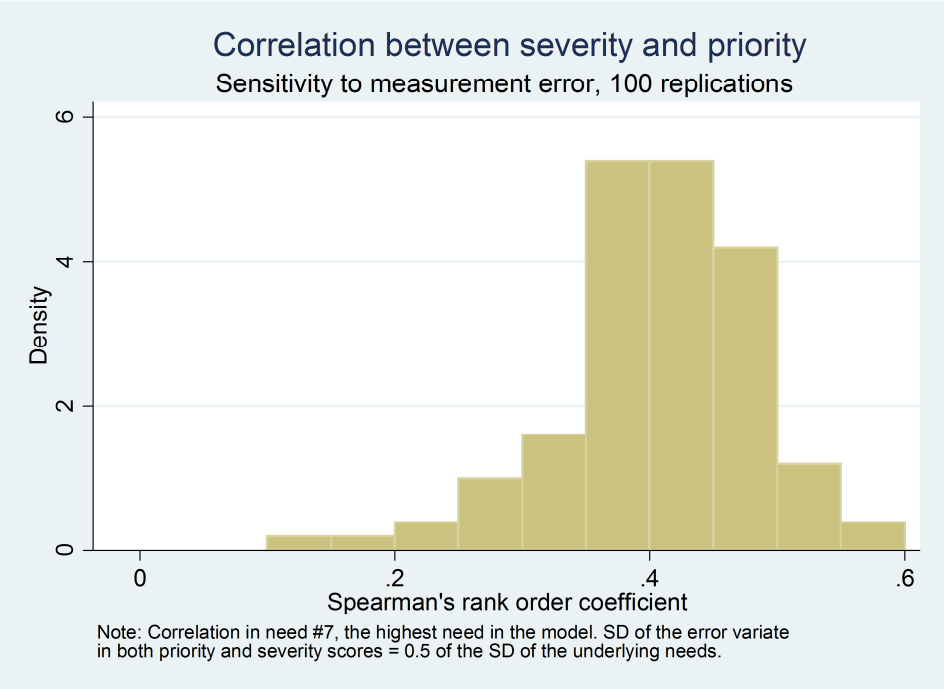
Severity score	Prioritiy score SD of the error variate				
SD of the error variate	0.00	0.25	0.50	0.75	1.00
0.00	0.50	0.50	0.46	0.40	0.35
0.25	0.49	0.49	0.44	0.40	0.34
0.50	0.45	0.46	0.41	0.36	0.31
0.75	0.41	0.39	0.37	0.32	0.28
1.00	0.36	0.35	0.32	0.28	0.24

From 0.50 for no error, the mean coefficient drops to 0.41 for 0.5 SD errors in both scores, or roughly a 20 percent loss. It drops to 0.24 at the extreme error levels considered, a loss of more than 50 percent. The values are almost symmetrical around the diagonal, indicating that equal measurement error in either score inflicts similar loss.

However, these losses are mean losses, over "100 different simulated worlds", if you will. This does not mean that in the one real world, correlations between severity and probability scores can be trusted without additional knowledge. The histogram below shows their variablility. A "real" correlation of 0.50 can almost disappear under error (as low as 0.15); rarely is it overestimated. In this simulation, the coefficient stayed between 0.26 and 0.52 in 90 percent of the replications.

[15] Which is one of the reasons why the Borda interpretation with legitimate means should be used whenever defensible.

Figure 8: Correlation between severity and priority scores under simulated measurement error



Practically, lack of observed correlation between the two measures of need does not imply that in reality they do not grow or diminish together. One needs additional knowledge to decide that. Notably, spatial associations - visualized in maps or estimated in some kind of regression model - will help illuminate intersectoral connections. Other additional comparators might be tried out from among the sets of most urgent problems (or "key issues") identified in each sector - provided these problems are sufficiently distinct semantically and pragmatically.

In sum, readers may wish to retain these points:

- The severity and priority scores for a given sector and given unit are statistically independent from each other. The severity scores are independent between sectors. The priority scores are not independent because they express rankings of sectors vis-à-vis each other. Thus, practically, it is ok to correlate the severity scores between any two sectors, but not the priority scores[16].
- As a mark of validity and reliability, the two scores for a given sector should be positively correlated. Except for the trivial case of considering two sectors only, the expected strength of the correlation will likely be modest for the sectors with the highest unmet needs, and weak for those of lesser needs.
- Under measurement error, these correlations are likely to be even weaker. This means the associations between severity and priority should be studied, but that

[16] Unless one goes into exotic models for compositional data (Thió-Henestrosa and Martín-Fernández 2005).

there is no reason to panic over validity or reliability as long as the scores exhibit at least a weak positive correlation. If GIS help is at hand, reading maps of severity and priority scores side by side will be helpful to deepen the understanding of where, and possibly why, they are aligned, and where not.

Prioritization by area and by sector

Severity and priority scores are computed in order to make comparisons among entities of key interest to the assessment - sectors (need domains) and locations (assessed units). Depending on how concepts and data formats are connected, the ambition is to extend the comparisons to regions (sets of locations) and to social groups (using demographic data and assumptions).

Two obstacles arise against naïve comparisons:

- First, assessed units differ in *population size*, sometimes by magnitudes. Giving all of the units - communities, camps, subdistricts, etc. - the same weight does not seem fair or conducive to good response planning. Thus weighting issues must be thought through, including the influence of outliers and the robustness of findings to weighting.

- Second, severity scores, as used so far, are *ordinal variables*. Their legitimate statistics include the rank-based percentiles (particularly the median), maxima and minima as well as the counting-based modes and frequencies (e.g., the proportion of units with health severity at levels 4 and 5). But not the mean. Population-weighted medians of severity scores are legitimate, but the sum of products from multiplying the score by population is not. This latter constraint particularly limits comparisons between locations.

Priority scores rank sectors. Locations cannot be compared over all sectors using these scores since the sums should be equal everywhere. They can be used to compare sectors, though. For this, the view of priority scores as interval-level constructs is defensible, as long as the assumption holds that the respondents understood the format. However, the Borda count treatment is of a so-called "modified" kind because lesser priorities are collectively assigned zero. This does not mean that there are no unmet needs in these sectors. Since they remain unranked, the scale does not have a meaningful zero point. The modified Borda count permits comparisons between units, not only by rank, but even as differences. The error risks in population-weighting priority scores are difficult to assess, but they are probably no greater than those caused by the choice of the number of items ranked in the unweighted modified Borda count[17].

[17] Consider a small 7-sector situation, with sectors X, Y, Z1 … Z5, and with only two communities, A and B. Let us assume that in community A, if informants are asked to rank all sectors, they assign X first rank, and Y third. In B, they assign X seventh, and Y third. In the *full* Borda count model, the ranks are translated into the scores, and the two sectors are compared, as follows: (7 + 1 = 8 < 5 + 5 = 10) = **(X < Y)**. If the priority choices are limited to three (a variant of the *modified* Borda count), then (3 + 0 = 3 > 1 + 1 = 2) =
[continued next page]

The challenges thus appear greater with severity than with priority. We will discuss some met in the use of unweighted severity scores as well as of the population-weighted variety.

Unweighted severity scores

For easy visualization, we assume ten locations assessed in five sectors, on a severity scale from 1 to 5. We first present two extreme cases. One is characterized by perfect dominance. The other demonstrates that complete individual differentiation can go hand in hand with complete lack of differentiation at the aggregate level.

Figure 9: Severity scores with perfect dominance pattern

Dominance

Location	Sectors Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Location median
1	1	1	2	2	3	2
2	1	1	2	3	3	2
3	1	2	2	3	3	2
4	1	2	3	3	4	3
5	2	2	3	3	4	3
6	2	3	3	4	4	3
7	2	3	3	4	5	3
8	3	3	4	4	5	4
9	3	3	4	5	5	4
10	3	4	4	5	5	4
Sector median	2	2.5	3	3.5	4	

As the staircase coloring of the severity score table makes easy to recognize, every subsequent location, moving from #1 to #10, dominates all preceding locations, and so does every one among the five sectors regarding all preceding sectors. In these ideal circumstances, inference to which of any two locations is more severely affected is straightforward. It is even stronger than reliance on the median score across sectors. Thus location # 2 is more severely affected than #1, despite equal median scores of 2, because in sector #4 it was scored 3, while location #1 has a score of 2 only. The comparison of sectors is even easier because their median scores over the locations are all different.

(X > Y), a preference reversal. The finding can be generalized to M communities just by duplicating the pattern for A and B as often as needed.

It is harder to see how different population distributions might affect the distribution of this kind of error. In defense of the modified Borda format, one must keep in mind that ranking is increasingly difficult and error-prone as the number of items to get an explicit rank (> 0) grows.

We proceed to the second extreme constellation:

Figure 10: Unit-level differentiation vs. aggregate uniformity

Maximum individual differentiation with aggregate uniformity

	Sectors					Location
	Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	median
Location						
1	1	2	3	4	5	3
2	1	2	3	4	5	3
3	2	3	4	5	1	3
4	2	3	4	5	1	3
5	3	4	5	1	2	3
6	3	4	5	1	2	3
7	4	5	1	2	3	3
8	4	5	1	2	3	3
9	5	1	2	3	4	3
10	5	1	2	3	4	3
Sector median	3	3	3	3	3	

In this contrived example, the scores - within each sector as well as within each location - are differentiated to the maximum. However, they perfectly equalize in the aggregate.

Those scenarios are extreme idealizations. We offer another contrived example, one in which the arrangement deviates from the perfect dominance. It demonstrates that the order of sectors and locations can differ, depending on which statistic of the severity score we apply in comparisons.

Figure 11: Pattern with high-severity clusters at opposing corners

Disturbed dominance

Location	Sectors Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Location median
1	5	5	4	4	3	4
2	5	5	4	3	3	4
3	5	5	4	3	3	4
4	1	2	3	3	4	3
5	2	2	3	3	4	3
6	2	3	3	4	4	3
7	2	3	3	4	5	3
8	3	3	4	4	5	4
9	3	3	4	5	5	4
10	3	4	4	5	5	4
Sector median	3	3	4	4	4	

The red border indicates the three locations for which the scores were altered from the perfect dominance scenario.

Relying on the median, the severity is greater in sectors #3 - 5 (median score = 4) than in #1 - 2 (median = 3). If we base the comparison on the number of locations with the maximum score (5, meaning "many people are already dying as a result of shortages in this sector" in the J-RANS II), a different conclusion arises. Among the sectors with median score = 4, #5 has four such locations, and #4 has two. However, now the sectors with lower median scores, #1 and 2, matter more than #4 - they each count three communities with severity at level five. Remember that both approaches - using the median score and the frequencies of values in a range of interest - are legitimate comparators.

In practice, such a situation may be exceptional, conceivably occurring in assessments that cover multiple disasters of different nature (e.g., Bolivia). But it gives us a hint that we might want to evaluate the robustness of findings by considering more than one statistic of the severity score, and in any event to be sure to compare severity and priority.

Population-weighted severity scores

The use of population-weighted median scores, over all sample locations, to characterize *sectors* by severity is unproblematic, except for what we just said about testing robustness[18].

[18] There is the very real technical difficulty that MS Excel has no built-in routine for calculating weighted medians. Suggested solutions on Web-based help sites all seem unpractical for non-integer weights (which *[continued next page]*

Comparing *locations* by looking at their severity scores across sectors and by taking into account their populations is difficult and in many constellations outright impossible. We therefore devote the next section to the difficulty with comparisons among a small number of communities that were assessed in the J-RANS II.

Comparisons of individual locations

The question is: How can locations (e.g. subdistricts) be compared regarding their unmet needs, considering their populations and their severity scores in the various sectors? For example, should a location with a level-4 ("many people will die") score in just one sector be considered less needy than another that scores 4 in three sectors?

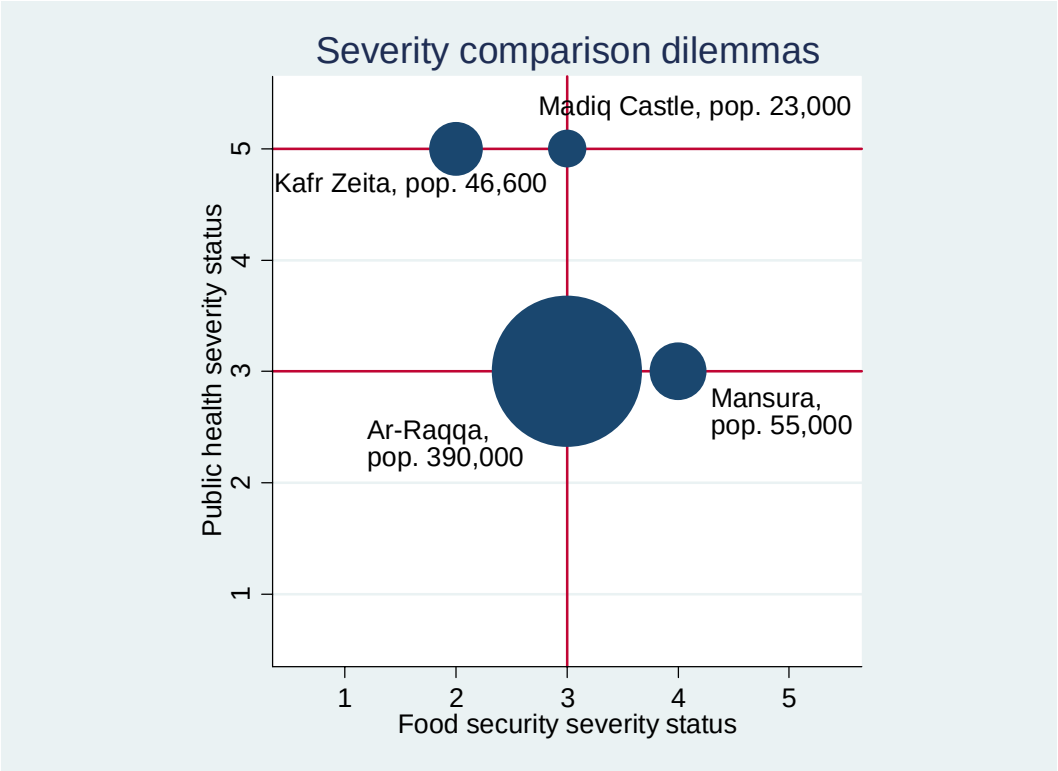
The challenge is that there is no ready measure that combines population size and severity, and even less so across sectors. To illustrate this, we made a didactic choice of four subdistricts in which these criteria are plainly in conflict with each other. For simplicity, we consider food and health needs only. Al-Raqqa has a population multiple times larger than that of any of the other three subdistricts in this chart, but in both sectors, its score is 3 ("many are suffering").

Since we cannot meaningfully multiply population and severity score, Ar-Raqqa and Mansura cannot be compared on the basis of these numbers only. The simple use of the maximum is not very convincing - it is believed that an unknown number of people in Mansura "will soon die" whereas nobody has said so about those of Ar-Raqqa. But "suffering" covers a wide spectrum of conditions, and with a population seven times larger than Mansura's, Ar-Raqqa may have a significant number of people whose sufferings, just shy of dying, is so elevated that, could we put a figure on it, "the total amount of their suffering" may well be higher than Mansura's, including the people vowed to death. It is simply not possible to judge.

---

do appear as soon as we normalize populations to sum to the number of observations). For readers needing to compute weighted medians, an intuitive workaround is to place, in a new worksheet, a copy of the weighting variable (most often the population) and the variable to be weighted side by side, sort on the latter ascendingly, calculate normalized weights (= weights summing to the number of observations), display their running sum in the next column, and the IF-formula *=IF(RC[-1] > (N/2),1,0)* [where N is the number of observations] down the next, and an adjustment formula for odd numbers of observations in the next. Name the ranges that hold the IF-formulas and the adjustment values. Finally, in the cell to receive the weighted median, collect its value with a lookup formula like *=INDEX(AdjustedValueRange, MATCH(1,IfFormulaRange,0),1).* If the author survives the protests that this reckless description is bound to provoke, he may try to write a user-defined function to which the scores and population weights (or other types of weights when appropriate) can simply be passed as named ranges.

Figure 12: Some constellations of unit-level severity comparisons



Not all comparisons are infeasible. Some can be argued on common sense and (assumedly) shared ethical convictions. Take Madiq Castle and Mansura. Their populations are not that much different, about 1 : 2.4. Mansura is worse off in terms of food security, but only up to level 4. Madiq Castle, however, is at level 5 in terms of health needs; there are many people "dying now". Under these circumstances, it may be plausible to say that, for the time being, the situation of Madiq Castle is worse than that of Mansura. However, this is without considering additional information, such as on needs in other sectors. Looking at all the information we may again conclude that 1. the opposite should be believed, 2. the comparison is impossible.

The discussion of other pairs of subdistricts in the chart does not contribute much more insight. Take, as a last example, Kafr Zeita and Madiq Castle, both with acute health problems from which many people are dying now. Kafr Zeita's population is double of Madiq Castle's. Its food problems warrant monitoring (level 2), but are not (yet) causing widespread suffering, as different from Madiq Castle (level 3). However, without knowing more about the death rates in the two areas, it is impossible to weight the severity of their situations against each other. Plus, as response planners will be quick to add, the comparison is academic: the people in Mansura need access to food most urgently; in Madiq Castle, health care support will reduce suffering most effectively.

[Sidebar:] What determines severity scores?

What could one learn from additional information? For one thing, it might be interesting to test what the local conditions are that cause assessment teams to rate areas as more or less severely affected in the different sectors. For an illustration, we turn again to the J-RANS II dataset. We limit ourselves to the two sectors health and food security and estimate the effects of

- Population
- High vs. low conflict intensity
- IDP rate in the current population

on the two severity scores. We use the magnitude (=log10) of the current population and control also for the possibility that conditions inside a governorate may be more similar than across governorates (clustering).

Without a detailed discussion of the statistical model[19], we summarize some of the outcomes:

1. *severity in health* tends to be higher in the more populous subdistricts and in those with higher conflict intensity. It is not significantly affected by the rate of displaced persons to the current population.

2. The *food severity* score does not respond to population size or conflict intensity. Though very weakly significant, it does tend to go up with the rate of displaced persons.

3. There is an indication that other, unobserved, factors are pushing the health and food severity in the same direction. Their factors could be contextual (the levels of unmet health and food needs grow side by side) or correlated measurement errors (some key informants exaggerate or underplay both needs) or both.

This does not make area comparisons on severity scores any easier, but it indicates that the severity of some unmet needs (health) is more strongly associated with certain characteristics of the areas than other needs (food). If the severity scales discriminated better, one would expect that the effects of the displacement rates were significant.

Naively, one might speculate that the health needs of areas with higher displacement rates might be relatively lower because people came to these areas for the better protection that they offered. Conversely, the food needs might be higher. But as, P. Chataigner of ACAPS observed (personal communication), *the health services are similarly compromised all over the country. Yet, it is to the low-conflict intensity areas that displaced persons flock, draining health care resources more heavily than in high-conflict intensity areas. Plus, injured persons are evacuated for treatment to low conflict intensity areas, further adding to the strain. Commonly, LCI areas receive less health sector assistance that HCI areas. As a result, we find that the situation is worse in the LCI areas".*

All this goes to say that local context drives the severity judgments, but that relationships between context and scoring are complex and in part obscured by unmeasured factors and by error in the measured ones (and by the biases introduced by purposive sampling). This insight is not particularly profound, except to suggest, again, that when we want to compare sectors or locations, more information than just the severity scores should be taken into account.

[19] Technically a bivariate ordered probit regression.

Towards better measurement

This section offers some cautious recommendations. Some may be productive in most assessment situations; others may be feasible only in particular contexts and conditions.

Use more than one measure: The J-RANS II experience supports the case that more than one measure should be employed in gauging the order of unmet needs. As in the Aleppo assessment earlier, there was a fair amount of agreement between the severity and priority, enough to confirm the sector with the highest unmet needs.

In Aleppo, the severity and priority scoring itself sufficed to find areas of agreement between the two measures. They agreed that health needs were high, and WASH needs less important. They disagreed on food, nutrition and shelter needs.

In the J-RANS II case, the differentiation of the severity scores was poor. The assessment team correctly switched to another statistic in order to express priority needs. It calculated the populations at risk (level 3 of the severity scale) and at high risk (levels 4 and 5) in each sector. It did so, however, only with the severity, not with the priority scores. When calculating populations at risk on both measures (on the priority side, we calculated populations of subdistricts in which the sector was of the first or second priority), agreement follows a similar pattern.

In Yemen, the correlations between severity and priority were weakly positive. But the instrument was not the same format as in Syria. The severity scale had four levels only; three, not five, priority sectors were elicited; the sector priorities were the communities' own while the severity scores were assigned by the assessment teams.

The two measures - if this experience is any indication - seem to agree at the extremes, but disagree on sector importance in the middle zone of unmet needs.

Revise the severity scale: The lack of differentiation in the data produced by this tool was one of the most serious limitations on the J-RANS II. This is not an isolated shortcoming of this assessment; it is common problem in survey research using ordinal scales. The elaboration and testing, in the next assessments or in a more continuously working humanitarian monitoring system, of a scale that differentiates better is very important.

Among the various improvements and alternatives discussed, the seven-level modification:

Table 22: Possible seven-level severity scale

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

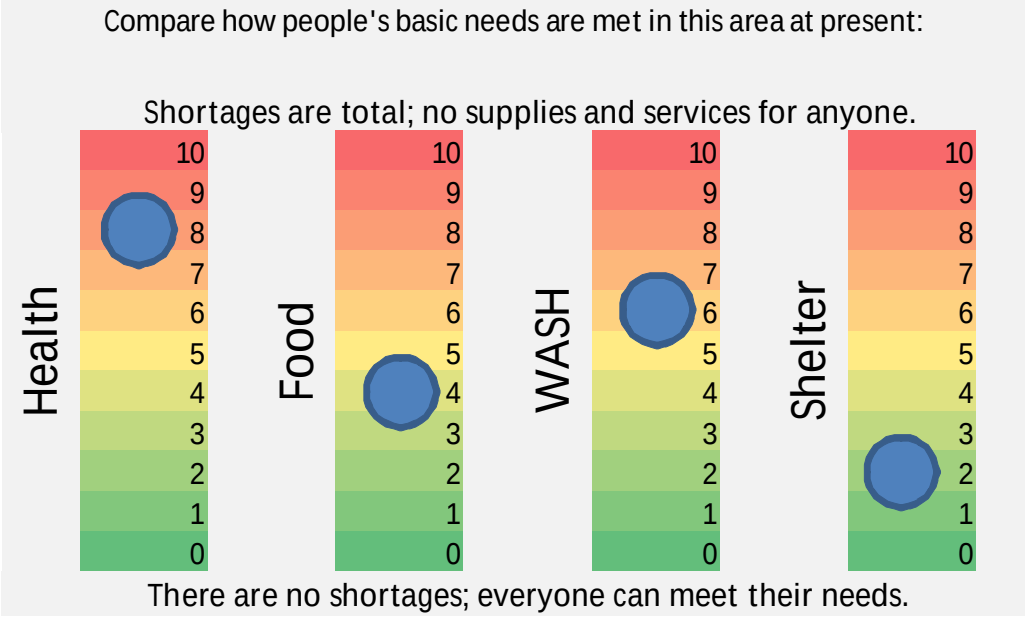
may be expected to work somewhat better than the current instrument. Its adoption should not be too risky because it elaborates on an already familiar tool. Visuals - the text of the scale filling a full A4-page - should be helpful.

Use fewer options in the priority scale: Regarding the sector priority scale, we should question whether key informants can meaningfully rank as many as five sectors. Would three priority sectors be more credible? Asking for three will plausibly save time, compared to five. It might be illuminating to ask a supplementary question of a qualitative kind: *"Why is this sector your first .. second .. third priority?"* or even with a comparative stimulus: *"You gave sector X as your first, and sector Y as your second priority. What makes you choose X as the first?"* The qualitative response would not have to be complete - in fact, cannot be complete -, but it will help assess the validity of the choices and even make for informative cross-referencing with severity ratings and most urgently required interventions.

Consider using a second severity measure: Besides the severity scale (in which all levels have an explicit, verbal meaning) and the priority scale, a ladder-type tool, with only the two endpoints clearly interpreted, should be developed as an additional measure. Where exactly to introduce it in the sequence of key informant activities or in the enumerators' synthesis report, is a question of research tactics. It certainly must be done in a way that does not make the conversations tiring or redundant.

For this, good visuals will be needed, and the concept that choices on the ladder express has to be communicated clearly. This figure suggests a fairly large (at least A4-size, ideally poster size) diagram placed in front of key informants, who are given tokens (blue chips here) to place on the color scales symbolizing the degree of current deprivation. Only the endpoints are defined, and only by the one concept of "shortage", reinforced with "no supplies and services for anyone", respectively at the other extreme: "everyone can meet their needs".

Figure 13: Notional template for a severity ladder visual



Since color printing may not be available at field offices, ACAPS may need to keep a stock of printed generic templates that will be labeled ad hoc in the assessment language.

Such a measure should be tested carefully and should be *supplementary* to the fully interpreted severity scale. In no way should the assessment designers gamble success on an untested severity ladder replacing the familiar scale.

Express severity with population figures: It was a creative step for the J-RANS II team to express severity by way of populations given certain levels of severity in four sectors (the often referred-to Figure 14, page 23, of the report). The re-interpretation as populations "at risk" and "at acute risk" also dispensed the report from the difficulty of defending the severity measure[20].

Such statistics are helpful, notably because assessment consumers easily understand them. The statement that "two million people are at acute risk for lack of health care" is more informative and didactic than the academic "the population-weighted median of the severity score in the health sector is 3". This way of presenting findings should continue.

In the internal analysis by the team, however, some controls are needed. The at-risk population distribution based on severity scores should be compared with the corresponding figures based on the priority score (see Table 5 on page 13). Also, if time and skills allow, the robustness should be tested through simulation, particularly for the impact of measurement error in population estimates.

[20] The numeric aspect of the scale is explained in the caption to a combined-needs map on page 6 of the J-RANS II report; some of the level meanings (such as "many people will soon die") are used in the sectoral priorities chapter; and the full interpretation of the generic severity scale is given on page 21.

Revise the elicitation of "serious problems" by sector: the J-RANS II elicited response to a pre-defined list of problems in each sector, with the option for respondents to nominate other problems (duly recorded in a text field). Some of the lists are long (10 pre-defined problems regarding food, eight regarding nutrition, etc.). Enumerators and key informants were instructed to select a maximum of five. Measurement was binary: "Yes, we have such a problem" or "No, we don't".

For the reasons explained in the section "Within-sector" problems on page 22, this format militates against valid data and valid analysis. At least three facets should be revised:

- The lists should be shorter.
- The problems need to be formulated such that they a more clearly distinct.
- The five-problem limit should be lifted.

In addition, the elicitation of "other" problems should be more forceful and encouraging. Why not simply demand some problems that are not yet listed? *"We are sure that in addition to the problems just now recorded, there are others, not yet found in this list. Please tell us what you feel are other urgent problems in this sector?"* Standardized interviewer stimulus is important for this purpose, and needs to be trained. Post-coding will be necessary. In the J-RANS II, the use of the "other" option revealed that in a number of subdistricts leishmaniasis (Wikipedia 2013b) had become a health problem. This suggests that efforts to stimulate non-trivial problem reports are worthwhile.

Shorter lists of problems, in clearly distinct formulations, also have a chance to offer items that can form mini-severity scales in their own right. For this to work, the items have to be statistically independent, and not constrained by a choice maximum.

ACAPS may want to consider devising such potential scales, sector-wise, in the shape of sets of generic questions drawn from its assessment question bank and from assessment tools of other agencies. The sets will have to be shortened, and the questions adapted to local context, whenever a new questionnaire is to be built.

Make greater use of highly educated enumerators: Many among the enumerators employed by the J-RANS were professionals or otherwise well educated individuals. Depending on the time table for design, training, data collection, debriefing, data entry and analysis, there is a potential to differentiate between the information to be elicited from key informants (which the enumerators record) and that which the enumerators can supply from outside the key informant contacts. Many sources could come into play: personal experience and familiarity, mobile phone calls to residents and displaced persons whom the enumerator trusts, browsing through lists and documents while in the area being assessed. Also, at the end of the stay in the area assessed, enumerators could fill rating and ranking type scales that are meant for them only, different from those presented to key informants.

We understand that several enumerators conducted multiple interviews in some or all of the subdistricts that they visited. They filled out a separate questionnaire in most of those interviews and later handed in a synoptic version that averaged the information for the subdistrict. The lower-level information was not used in data entry - there are no records of units lower than the subdistrict[21].

To the extent that within-subdistrict conditions vary greatly and in ways important for the assessment, such lower-level information might be put to productive use. A partial data entry might be justified, picking out, for that level, only the population, severity and priority measures. This can be justified by the generic possibility that in multi-level setups (village - electoral ward - subdistrict - district, etc.) it is hard to predict how the variability of items of interest is distributed over the various levels. And in a conflict perspective, conditions may vary greatly at local levels, including those aspects that are key to the assessment.

Conclusion

The rating and ranking of unmet needs by sectors is a difficult challenge in all assessments, whatever their specific formats. Categorical scales are meant to make the intensity of needs comparable across sectors. There are many ways of constructing scales, but apparently few to validate them, given the inaccessibility of most needs to one-dimensional summary. In this situation, it is reasonable to work with more than one needs measure per sector.

ACAPS has developed assessment templates that work with two measures - the severity and priority scales. There have been difficulties; in both the Yemen and the Syria assessments, the severity scales did not differentiate enough. This seems to be the point that needs repairs most decidedly.

It is easy to find out, after the event, that a scale did not perform well, but tricky to pinpoint why that happened. In the J-RANS II setup, the data reflect the enumerators' syntheses of what they learned in encounters with one, and often with several, groups of key informants in a subdistrict. We do not know how the key informants processed the questions, some of which made high cognitive demands. Low differentiation in the severity scale response may be due to any of three factors: the reality on the ground was such that, correctly, one particular score was given frequently; key informants, not understanding the question, settled for a middling level that seemed acceptable; there was significant variability within subdistricts, and the high frequency of the middle category is the result of enumerators averaging across key informants.

All this boils down to being more conservative in our beliefs of what key informants can handle, and to be bolder with regards to what highly educated enumerators like Syria's can manage. With more assessments in that country likely coming up, and with a view to other countries as well, one hopes that cumulative learning takes place. With the

[21] With one possible exception where separate records were created for the rural and urban parts of a subdistrict.

experience of three assessments in Syria and with several in other countries, the instruments to measure the severity and priority of needs can to a fair degree be prepared in advance. Some local assembly will be required, as the saying goes, but the tools will come together faster and will work better.

References

ACAPS (2011). Joint Rapid Assessment of the Northern Governorates of Yemen [9 October 2011]. Sana'a, Assessment Capacities Project (ACAPS), in collaboration with ADRA Yemen, CARE International, Save the Children, OXFAM, and Islamic Relief. Prepared for CARE International in Yemen.

ACU (2013). Joint Rapid Assessment of Northern Syria. Final Report [17 February 2013], Assistance Coordination Unit (ACU), supported by ECHO, DFID and OFDA.

Alwin, D. F. and J. A. Krosnick (1985). "The measurement of values in surveys: A comparison of ratings and rankings." Public Opinion Quarterly **49**(4): 535-552.

AWG (2013a). Joint Rapid Assessment of Northern Syria - Aleppo City Assessment [28 March 2013], Assessment Working Group for Northern Syria.

AWG (2013b). Joint Rapid Assessment of Northern Syria II. Final Report [22 May 2013], Assessment Working Group for Northern Syria.

Benini, A. (2011a). Data analysis in needs assessments [6 November 2011]. Geneva, Assessment Capacity Project.

Benini, A. (2011b). A template for managing data in needs assessments, centered on sites, sectors, problems, and severity of needs [21 October 2011]. Geneva, Assessment Capacity Project.

Coates, J., E. A. Frongillo, et al. (2006). "Commonalities in the Experience of Household Food Insecurity across Cultures: What Are Measures Missing?" The Journal of Nutrition **136**(5): 1438S-1448S.

Cojocaru, A. and M. F. Diagne (2013). How reliable and consistent are subjective measures of welfare in Europe and Central Asia? Evidence from the second life in transition survey [Policy Research Working Paper No. 6359]. Washington DC, The World Bank.

De Chiusole, D. and L. Stefanutti (2011). "Rating, ranking, or both? A joint application of two probabilistic models for the measurement of values " TPM-Testing, Psychometrics, Methodology in Applied Psychology **18** (1): 49-60.

King, G., C. J. L. Murray, et al. (2004). "Enhancing the validity and cross-cultural comparability of measurement in survey research (vol 97, pg 567, 2003)." American Political Science Review **98**(1): 191-207.

Klein, M., H. Dülmer, et al. (2004). "Response sets in the measurement of values: A comparison of rating and ranking procedures." International Journal of Public Opinion Research **16**(4): 474-483.

Kolenikov, S. and G. Angeles. (2004)."The Use of Discrete Data in Principal Component Analysis With Applications to Socio-Economic Indices. CPC/MEASURE Working paper No. WP-04-85." Retrieved 21 January 2005, from https://www.cpc.unc.edu/measure/publications/pdf/wp-04-85.pdf.

Langville, A. N. and C. C. D. Meyer (2012). Who's N° 1?: The Science of Rating and Ranking, Princeton University Press.

Maio, G. R., N. J. Roese, et al. (1996). "Rankings, Ratings, and the Measurement of Values: Evidence for the Superior Validity of Ratings." Basic and Applied Social Psychology **18**(2): 171-181.

NIH (2003-07). Pain Intensity Instruments. Bethesda, Maryland, National Institutes of Health - Warren Grant Magnuson Clinical Center.

Noble, M., G. Smith, et al. (2003). "Scottish indices of deprivation 2003." Edinburgh: Scottish Executive.

Ravallion, M. (2012). Poor, or Just Feeling Poor? On Using Subjective Data in Measuring Poverty [Policy Research working paper # 5968]. Washington DC, The World Bank.

Roszkowski, M. J. and S. Spreat (2012). "You Name It: Comparing Holistic and Analytical Rating Methods of Eliciting Preferences in Naming an Online Program Using Ranks as a Concurrent Validity Criterion." International Journal of Technology and Educational Marketing (IJTEM) **2**(1): 59-79.

Sayre, S. (2006). Using video-elicitation to research sensitive topics: understanding the purchase process following natural disaster. Handbook of Qualitative Research Methods in Marketing. 230-.

Thió-Henestrosa, S. and J. Martín-Fernández (2005). "Dealing with compositional data: the freeware CoDaPack." Mathematical Geology **37**(7): 773-793.

WHO and King's College London (2011). The Humanitarian Emergency Settings Perceived Needs Scale (HESPER): Manual with Scale. Geneva, World Health Organization.

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. (2013a)."Arrow's impossibility theorem." Retrieved 20 June 2013, from http://en.wikipedia.org/wiki/Arrow%27s_impossibility_theorem.

Wikipedia. (2013b)."Leishmaniasis." Retrieved 26 July 2013, from http://en.wikipedia.org/wiki/Leishmaniasis.

Wikipedia. (2013c)."Rasch model." Retrieved 7 August 2013, from http://en.wikipedia.org/wiki/Rasch_scale.

Wikipedia. (2013d)."Vignette (psychology)." Retrieved 18 June 2013, from http://en.wikipedia.org/wiki/Vignette_(psychology).

Appendix: Simulation code (Stata do-file)

```
*******************************************************************************
* SIMULATION OF THE AGREEMENT BETWEEN TWO NEEDS MEASURES IN RAPID NEEDS ASSESSMENTS:  *
* 1. SEVERITY (ordinal; independent for each need variable) and                       *
* 2. PRIORITY (interval-level under Borda count assumptions; relative to other needs) *
* under various levels of measurement error.                                          *
*                                                                                     *
* Aldo Benini for the Assessment Capacity Project (ACAPS), Geneva, Switzerland        *
* as part of a review of the Syria J-RANS II Assessment.                              *
*                                                                                     *
* Version: 25 June 2013. Highlighted areas: User to define working directory          *
*******************************************************************************

set more off

*************************************************************************
* PART 1: SIMULATED NEEDS VARIABLES, INITIALLY WITHOUT MEASUREMENT ERROR *
*************************************************************************

* Generating the random variates for needs, the derived needs scores and (Borda-scored)
priorities
* Working with 7 different need sectors; no measurement error at this stage.
* The needs sectors are not substantively identified here.
*  In the J-RANS II they were: Health, food security, nutrition, WASH, shelter/NFI,
education, protection.
* However, severity scores were elicitied for five sectors only. This simulation is not
limited in this way.
* Assuming continuous underlying need variables, weakly correlated (0.4 between any two
of them), which is
* close to the median Spearman's rank correlation observed among sector severity scores
in the J-RANS II.

* Setting a working directory:
cd C:\...

* Correlation structure stored in:
use "C:\...\130623_1136AB_CorrelationStructure.dta", clear

mkmat need1-need7, matrix(Needscorr)
mkmat nmeans, matrix(Needmeans)
set obs 100
corr2data needreal1 - needreal7,  corr(Needscorr) seed(1002) means(Needmeans)
summ needreal*
corr needreal*
drop need1-need7 nmeans
matrix drop Needscorr
matrix drop Needmeans

save "C:\...\130623_1145AB_NeedsIn100CommunitiesSimul.dta", replace
* Save working copy for the error simulation:
save "C:\...\130623_1447AB_SeverityPriorityCorr_w_Error.dta", replace

* Create the categorical severity scores,
* between 1 (no shortages) and 5 (many people are already dying as a result of shortages
[in the sector in point])
forvalues i = 1(1)7 {
gen needscore`i' = needreal`i' if needreal`i'  ~=.
replace needscore`i' = round(needscore`i')
replace needscore`i' = 1 if  needscore`i' < 1
replace needscore`i' = 5 if  needscore`i' > 5 & needreal`i'  ~=.
}

* Create a rank variable for every underlying need variable, with 7 being the highest
need
rowranks needreal1- needreal7, gen(needrank1-needrank7)
       * "rowranks" is not a standard command in STATA. It was written by Nicholas J. Cox
and advertised
       * in his "Speaking STATA: Rowwise", The Stata Journal, 2009, 9/1, 137-157. To
install it, type
       * "findit pr0046" in the command box (without the inverted commas) and go from
there.

* Correlations between each scored need and the ranks:
* [Of interest only because below we generate a Borda score limited to three priorities]
```

```
forvalues i = 1(1)7 {
di "Correlation for need `i' :"
spearman needscore`i' needrank`i'
}

* Generate Borda scores, but score only the highest three ranks, as 3, 2 and 1, and set
the others to zero:
forvalues i = 1(1)7 {
gen Borda3opt_`i' = needrank`i' - 4
replace Borda3opt_`i' = 0 if Borda3opt_`i' <0
}

***********************************************
* SUMMARY STATISTICS AND MEASURES OF AGREEMENT *
***********************************************
capture drop Needs67SevPriAgree
* In case this variable already exists.

* Summary statistics of interest:
tabstat needreal* needscore* Borda3opt_*, statistics( mean p50 min max sd) c(s)
* The results - in this particular simulation setup - show concordance between
* severity and (Borda-scored) priority for the two highest needs, needs #6 and 7.
* This is based on the comparison of medians of the ordinal severity scores and
* the Borda scores (for the latter, it is indifferent whether the ordinal [medians] or
cardinal
* [means] interpretations are used).

* Therefore we look at the degree of concordance over all cases:

* 1. by means of the correlations between scored needs and the truncated Borda score:
       forvalues i = 1(1)7 {
       di "Correlation for need `i' :"
       spearman needscore`i' Borda3opt_`i'
       }
* 2. by the extent to Needs #6 and 7 are correspondingly rated in severity and priority
scores:
       gen byte Needs67SevPriAgree = ( needscore6 + needscore7 >= 8) *( Borda3opt_6+
Borda3opt_7 == 5) ///
        +  ( needscore6 + needscore7 < 8) *( Borda3opt_6 + Borda3opt_7 < 5) /* The "*"
and "+" outside the parentheses work as logical operators. */
       summ Needs67SevPriAgree
* [The agreement is not 100% because: 1. rounding and truncation in severity scores, 2.
interference by other needs variables in the Borda scores.]

save "C:\...\130623_1447AB_SeverityPriorityCorr_w_Error.dta", replace

***********************************************************************
* COMBINED GRAPH, TO DEMONSTRATE CONNECTION BETW. SEVERITY AND PRIORITY *
* Exemplified with highest need (need # 7) and lowest (need # 1)        *
* Produces raw graph, to be edited for titles                           *
***********************************************************************

capture drop group1 group7 tag1 tag7 group1total group7total one
* In case these variables already exist.

gen byte one = 1

*Panel 1:
* Underlying needs, lowest vs. highest:
twoway scatter needreal7 needreal1, name(scatterreal, replace)
* graph save Graph "C:\...\ScatterNeedreal7vs1.gph", replace

* Panel 2:
* Bubble graph severity scores cross-tabulation
preserve
contract needscore7 needscore1
twoway scatter needscore7 needscore1 [aw=_freq], yscale(range(0.5 5.5)) ylabel(1(1)5)
xscale(range(0.5 5.5)) xlabel(1(1)5) name(scatterscores, replace)
* graph save scatterscores "C:\...\ScatterNeedscore7vs1.gph", replace
restore

* Panel 3:
* Bubble graph Borda scores cross-tabulation
preserve
contract Borda3opt_7 Borda3opt_1
twoway scatter Borda3opt_7 Borda3opt_1 [aw=_freq], yscale(range(-0.5 3.5)) xscale(range(-
0.5 3.5)) ylabel(0(1)3) xlabel(0(1)3) name(scatterBorda, replace)
*graph save scatterBorda "C:\...\ScatterBorda7vs1.gph", replace
restore
```

```
* Auxiliary variables needed to produce two Bubble graphs crossing severity and priority
(Borda) scores,
* one for need 1, one for need 7:
egen group7 = group( needscore7 Borda3opt_7)
egen tag7 = tag( group7)
egen group1 = group( needscore1 Borda3opt_1)
egen tag1 = tag( group1)
bysort group7: egen group7total = total(one)
bysort group1: egen group1total = total(one)

* Panel 4:
* Two combined Bubble graphs for severity vs. priority:
twoway scatter Borda3opt_7 needscore7 [aw = group7total] if tag7, yscale(range(-0.5 3.5))
xscale(range(0.5 5.5)) ylabel(0(1)3) xlabel(1(1)5) name(Bubble7, replace)
* graph save Bubble7 "C:\...\Bubble7.gph", replace
twoway scatter Borda3opt_1 needscore1 [aw = group1total] if tag1, yscale(range(-0.5 3.5))
xscale(range(0.5 5.5)) ylabel(0(1)3) xlabel(1(1)5) name(Bubble1, replace)
* graph save Bubble1 "C:\...\Bubble1.gph", replace
graph combine Bubble1 Bubble7, name(TwoBubbles, replace)
* graph save TwoBubbles "C:\...\TwoBubbles.gph", replace

* Combined raw graph:
graph combine scatterreal scatterscores scatterBorda TwoBubbles
* graph save Graph "C:\...\CombinedFourPanelsEdited.gph", replace

drop group7 - group1total
save "C:\...\130623_1447AB_SeverityPriorityCorr_w_Error.dta", replace

************************************************************
* PART 2: MEASUREMENT ERRORS IN SCORING AND PRIORITIZING  *
************************************************************

* PRELIMINARIES:

* Create an empty shell for the collection of simulation results at the bottom of the do
file:
        clear
        gen recno = _n
        forvalues i = 1(1)7 {
        gen NeedSc`i'Med = .
        gen PriorSc`i'Med = .
        gen PriorSc`i'Mean = .
        gen Corr`i' = .
        }
        gen Needs67_agree = .
        save CollectSimResults2, replace


* THE PROGRAM FOR THE SIMULATION PART:

* The program part, which the simulate command (below) calls to generate observations
with measurement error.
capture program drop SeverPriorCorrel /* capture ignores the error if there is no program
"SeverPriorCorrel" to drop */

* Create variables for the error-laden needs score base and needs priority base
use "C:\...\130623_1447AB_SeverityPriorityCorr_w_Error.dta", clear

        forvalues i = 1(1)7 {
        gen needscore`i'err = .
        gen priorbase`i'err = . /* The error-laden underlying needs variables on which to
compute priorities, then Borda scores. */
        gen Borda3opt_`i'err = .
        }

save "C:\...\130623_1447AB_SeverityPriorityCorr_w_Error.dta", replace

program SeverPriorCorrel, rclass
version 12

* Define the program's arguments:
args errormultsever errormultprior
* These are names for the error multiplication factors used in the formulas below. The
simulate command will pass values to them.
* Further below, in the simulation part, errormultsever will be represented by "k",
errormultprior by "j".

* Access the data file with the "observed" variables:
use "C:\...\130623_1447AB_SeverityPriorityCorr_w_Error.dta", clear

* Housekeeping stuff:
```

```
        capture drop severerr*
        capture drop priorerr*
        capture drop needrankerr*
        capture drop Borda3opt*err
        capture drop Needs67ErrAgree

* Generate error factors, and hence the severity and Borda scores with errors:
        forvalues i = 1(1)7 {

        * Create the severity scores with error:
        gen severerr`i' = rnormal() * 0.25 * `errormultsever' /* Error factor */
        * The simulation steps up errormultsever from 0 to 4; i.a.w. at the maximum,
        * the error component will have the same SD (= 0.25 * 4 = 1) as the simulated
needs.
        * This would make for rather stark mean absolute errors; the first two steps up
        * (up to half the SD of the simulated needs) should be more realistic.
        * [Same remark holds for the priorities with error below.]

        * ADDITIVE ERROR MODEL
        replace needscore`i'err = severerr`i' + needreal`i'
        replace needscore`i'err = round(needscore`i'err)
        replace needscore`i'err = 1 if  needscore`i'err < 1
        replace  needscore`i'err  =  5  if   needscore`i'err  >  5  &  needreal`i'   ~=.  /*  "&
needreal`i'  ~=." only for good manners. Not really needed. */


        * Create the bases with error for the Borda scoring:
        * [We assume that the scoring and priority setting errors are independent.]
        gen priorerr`i' = rnormal() * 0.25 * `errormultprior' /* Error factor; see remarks
above. */
        replace priorbase`i'err = priorerr`i' + needreal`i'
        }

        rowranks priorbase*err, gen(needrankerr1-needrankerr7)

        forvalues i = 1(1)7 {
        gen Borda3opt_`i'err = needrankerr`i' - 4
        replace Borda3opt_`i'err = 0 if Borda3opt_`i'err < 0
        }

*************************************************************************
* Calculate statistics on the error-laden severity and priority scores: *
*************************************************************************

* Medians of severity scores; medians and means of Borda scores:
        forvalues i = 1(1)7 {
        summ needscore`i'err, detail
        return scalar needsc`i'err_med = r(p50)
        summ Borda3opt_`i'err, detail
        return scalar Borda3opt_`i'err_med = r(p50)
        return scalar Borda3opt_`i'err_mean = r(mean)
        }

* Spearman's rank order correlations between the severity scores and the modified Borda
scores:
        forvalues i = 1(1)7 {
        spearman needscore`i'err Borda3opt_`i'err
        return scalar SpearmanCorr`i' = r(rho)
        }

* Agreement statistic for needs #6 and 7, to either be the highest on both severity and
priority, or neither:
        gen  byte  Needs67ErrAgree  =  (  needscore6err  +  needscore7err  >=  8)  *
( Borda3opt_6err + Borda3opt_7err == 5) ///
                + ( needscore6err + needscore7err < 8) * ( Borda3opt_6err+ Borda3opt_7err
< 5)
        summarize Needs67ErrAgree
        return scalar Needs67Agree = r(mean)


 end

* TESTING THE PROGRAM
* Unstar these three lines if you want to test the program:
* set seed 1111
* SeverPriorCorrel 1 1 /*Arbitrary test values for the two arguments. Because > 0, they
do cause observed and true to differ, as intended. */
* exit

*************************
* PART 3: THE SIMULATION  *
```

```
**************************

* The simulation part, in which forvalues augments the measurement error factor in steps
from 0 (no error) to 4
* the severity and modified Borda scores:

* Simulation command:

forvalues k = 0/4 {
	forvalues j = 0/4 {
	local seedi = 1234 + `k' + 10 * `j' /*Changes the random number seed at the start
of each simulation run as we augment the error mult. factor.*/
	set seed `seedi'
	simulate 		NeedSc1Med = r(needsc1err_med) ///
				NeedSc2Med = r(needsc2err_med) ///
				NeedSc3Med = r(needsc3err_med) ///
				NeedSc4Med = r(needsc4err_med) ///
				NeedSc5Med = r(needsc5err_med) ///
				NeedSc6Med = r(needsc6err_med) ///
				NeedSc7Med = r(needsc7err_med) ///
				PriorSc1Med = r(Borda3opt_1err_med) ///
				PriorSc2Med = r(Borda3opt_2err_med) ///
				PriorSc3Med = r(Borda3opt_3err_med) ///
				PriorSc4Med = r(Borda3opt_4err_med) ///
				PriorSc5Med = r(Borda3opt_5err_med) ///
				PriorSc6Med = r(Borda3opt_6err_med) ///
				PriorSc7Med = r(Borda3opt_7err_med) ///
				PriorSc1Mean = r(Borda3opt_1err_mean) ///
				PriorSc2Mean = r(Borda3opt_2err_mean) ///
				PriorSc3Mean = r(Borda3opt_3err_mean) ///
				PriorSc4Mean = r(Borda3opt_4err_mean) ///
				PriorSc5Mean = r(Borda3opt_5err_mean) ///
				PriorSc6Mean = r(Borda3opt_6err_mean) ///
				PriorSc7Mean = r(Borda3opt_7err_mean) ///
				Corr1 = r(SpearmanCorr1) ///
				Corr2 = r(SpearmanCorr2) ///
				Corr3 = r(SpearmanCorr3) ///
				Corr4 = r(SpearmanCorr4) ///
				Corr5 = r(SpearmanCorr5) ///
				Corr6 = r(SpearmanCorr6) ///
				Corr7 = r(SpearmanCorr7) ///
				Needs67_agree = r(Needs67Agree) ///
				, reps(100) nodots: SeverPriorCorrel `k' `j'
				/* Not very elegant: attempts to use forvalues all causes
errors */
	summarize
	tempfile results
		gen byte ScoreErrorFactor = `k'
		gen byte PriorErrorFactor = `j'
		save "`results'", replace
	use CollectSimResults2, clear
	append using "`results'"
	replace recno = _n
	save CollectSimResults2, replace

	}
}

**********************
* Tables of interest: *
**********************

* 1. Summary stats:
bysort ScoreErrorFactor PriorErrorFactor: summ NeedSc*Med PriorSc*Med PriorSc*Mean

* 2. Correlation between severity and priority scores - Robustness to measurement error:
* [Example: Highest need (need #7).]
table ScoreErrorFactor PriorErrorFactor, c(mean Corr7)

* 3. Agreement, as regards needs #6 and 7, between severity and priority rating - in
response to error levels:
table ScoreErrorFactor PriorErrorFactor, c(mean Needs67_agree)

* Housekeeping:
set more on
* Unstar "exit" if the variables with error of the last simulation run are to be kept.
* exit
use "C:\...\130623_1447AB_SeverityPriorityCorr_w_Error.dta", clear
capture drop Needs67SevPriAgree
capture drop needscore1err - Borda3opt_7err
save "C:\...\130623_1447AB_SeverityPriorityCorr_w_Error.dta", replace
```