Aldo Benini

A note for ACAPS

Composite measures

Their use in rapid needs assessments

Conceptual background and
technical guidance

Version 17 May 2012

Contents

Tables and Figures

Acknowledgement

Emese Csete, formerly with UNOCHA Cambodia, reviewed the first draft and provided a number of important corrections and suggestions. Kathleen Tierney, with the Natural Hazards Center, University of Colorado, helped with literature references.

Acronyms

ACAPS	Assessment Capacities Project
DEA	Data envelopment analysis
FAO	Food and Agricultural Organization
GFM	Global Focus Model
GIS	Geographical information system
GNP	Gross national product
HDI	Human Development Index
MADC	Multi-attribute decision making
OECD	Organization for Economic Co-operation and Development
PISA	Program for International Student Assessment
UNDP	United Nations Development Program
UNOCHA	United Nations Office for the Coordination of Humanitarian Affairs

Terminology

With over 260,000 hits in Google, "composite measure" opens the field for vague and divergent terminologies. The term is often used interchangeably with "composite indicator", "index", "total score" and "metric". While the latter three are at home in particular academic disciplines - indices, for example, are the oxygen of economists -, "composite indicator" is something of a contradiction in itself.

We speak of composite measures or simply "composite", plural: "composites". Sometimes, such as after an occurrence of "composite measure" in the text, we abbreviate to "measure". The variables of which the composite is formed are called "indicators" or "base indicators" or "variables". Terminological difficulties arise with multi-level composite measures; some or all of their indicators themselves are composites. For them, we use "subindices" (singular: subindex).

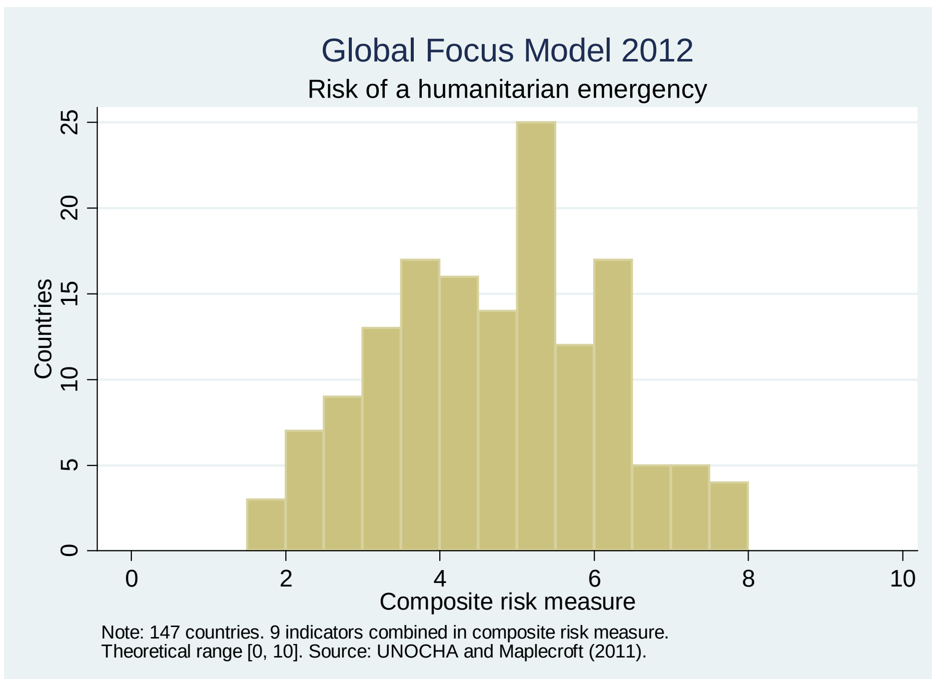
Summary

This note discusses composite measures in the context of rapid needs assessments.

Subject and motivation

A composite measure is the result of a mathematical function that maps several variables - its base indicators - onto a one-dimensional construct, such as an index number or an ordered category set. The Gross National Product and the country score on PISA student tests are examples. The credit rating for consumers - probably yes, but with algorithms that are not transparent to the concerned public!

Figure 1: An example of a composite measure from the humanitarian realm



Rapid needs assessments have several motivations for composite measures. Disaster impacts are correlated, and multiple impacts have cumulative effects on human suffering and on recovery chances. Hence the interest to express the total impact or the total unmet needs in a combined measure. Other motives include endeavors to impose order on information (so that, for example, maps can be colored by a ranked measure), to combine data from different sources (particularly when they provide different geographical coverage), and to balance measurement error.

Except for the need to capture correlated impacts, the case for composite measures in rapid needs assessments is weak. It is not clear what is being measured: current human suffering; current institutional disruption; subsequent recovery burdens. Validation against an external criterion is difficult - the criterion itself (e.g., excess mortality) is too narrow, or data are unavailable in useful time (e.g., a recovery index).

We are therefore skeptical of the feasibility and value of such constructs in rapid needs assessments. The tone is: "Avoid them, look for alternatives, but if you must use them, do this, don't do that."

Conceptual resources

We proceed as follows: The introduction situates the subject in the worldwide accountability revolution, in its various purposes within needs assessments, and in the basic constellation of pre- and post-disaster measurement. For theoretical orientation on the mechanics of composite measures, we rely on the OECD "Handbook on Constructing Composite Indicators". This, however, reflects a slower-paced, more analytic and data-rich environment than assessment teams will find after a disaster; we use the Handbook's checklists, but with different emphases and recommendations.

No less importantly, a branch of decision sciences called "multiple attribute decision making" has extensively investigated the operations that are useful and legitimate for composite measures; we used, and warmly recommend as a brief, accessible text, the introduction by Yoon and Hwang (1995).

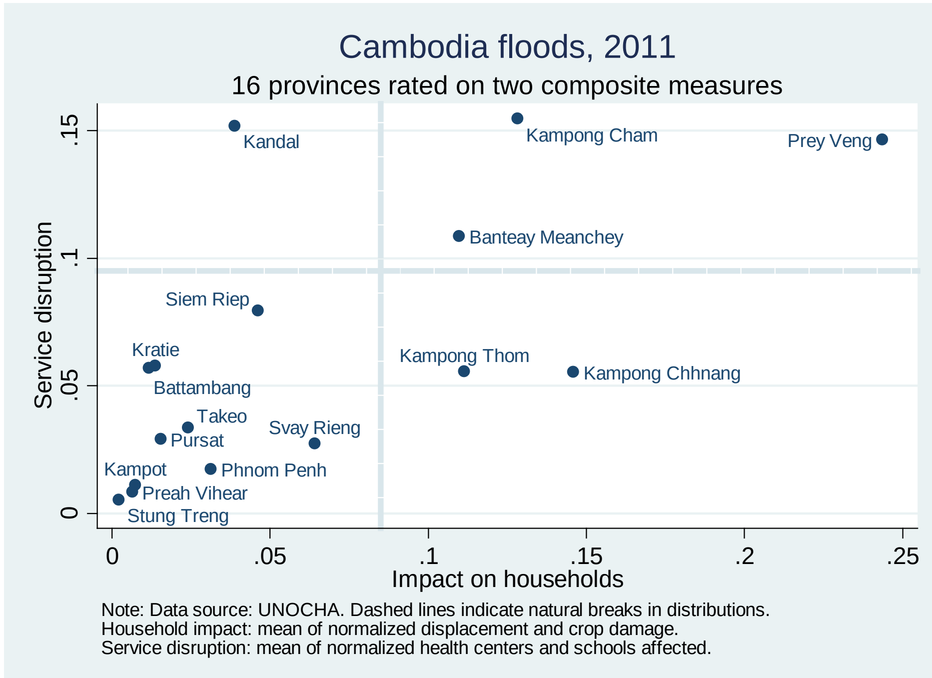
Examples from the humanitarian realm

For a more engaging access to the key aspects of our subject, we describe, discuss and selectively recalculate two applications from the humanitarian area. The reader may reproduce some of the calculations in the Excel workbook that comes with this note:

- The **Global Focus Model**, an UNOCHA initiative, produces annually updated risk estimates regarding humanitarian emergencies. Its 2012 data set is well suited to demonstrate a hierarchical indicator set-up, the robustness of the ensuing country ranks to methodological choices as well as some methods for rapidly X-raying the internal structure of the indicators.
- The composite measure that the UNOCHA analysts working on the **Cambodia floods of 2011** developed is our one example from a truly rapid assessment context. The design of the measure is simple and clear - and liable to be the target of a stinging criticism leveled at composite measures: that the weights assigned to the indicators are arbitrary.

We review, and advance some of our own, defenses against that charge and in the process develop an alternative to the one-tells-all composite: two substantively motivated subindices that together stake out an impact typology of provinces affected by the floods. The intent is captured in this figure.

Figure 2: Subindices - an alternative to one-tells-all composite measures



Step-by-step guidance

Following the review of those two practical applications, we develop rationales and step-by-step guidance for navigating the design and calculation of composite measures along the road taken by the OECD Handbook. We recommend methods and a degree of analytic depth that are compatible with 1. the pressures of conducting assessments after a sudden-onset disaster, 2. the use, by the assessment teams, of spreadsheet applications. It has to be admitted, however, that the testing of composite measures quickly runs into statistical complexities that make spreadsheets inefficient. In sidebars, we illustrate some analysis forms that fully-fledged statistical applications perform elegantly and rapidly.

This section covers:

1. Theoretical framework
2. Data selection
3. Imputation of missing data
4. Multivariate analysis
5. Normalization
6. Weighting and aggregation
7. Robustness and sensitivity
8. Back to the real data
9. Links to other variables
10. Presentation and visualization

We do not follow the Handbook's preference for so-called non-compensatory methods. Compensatory methods are those that offset the effect of an increase in one indicator by decrease in one or several others. While we demonstrate, in a sidebar using the Cambodia data, one of the many non-compensatory methods (in an Excel spreadsheet that may help users to start their own experiments), we hold that weights do express differences in importance between indicators of impact and need.

Graded recommendation

Our overall recommendation, in descending order of preference, is:

1. Form several subindices, not one all-encompassing composite. Each subindex must be substantively motivated (= its base indicators all speak to some plausible common concept). Visualize the diversity of impacts (or needs) in typologies, using scatterplots, Venn diagrams or summary tables.

2. If you must use one composite measure, use a compensatory method with unequal weights, each of them justified on policy grounds against the stated purpose of the measure.

3. If unequal weights are not feasible (because of lack of rationales or the risk of rejection by stakeholders), use unit weights (= equal weights).

4. If a compensatory method is not acceptable (because one of the indicators is supremely important), use a non-compensatory method. Excel users can implement Tversky's Lexicographic Semiorder Method.

If you are determined to form a composite measure, apply these tests

The decision to form a composite measure should be made on four criteria, in this order:

1. The measure serves a defined chief purpose.
2. It expresses one relevant concept and does so consistently.
3. It is feasible in terms of data acquisition as well as analysis.
4. It is agreeable to stakeholders.

We conclude with an outlook that reiterates our skepticism, but admits that new technologies, particularly in remote measurement and data transmission, may eventually help produce better validated composite measures for rapid needs assessments.

Recommended readings

For the reader anxious to familiarize with the subject in greater depth, the above-mentioned OECD Handbook (Nardo, Saisana et al. 2008) is a detailed, if sometimes irritating resource (several of their recommended methods have no readily accessible implementation). Yoon and Hwang (1995), although somewhat outdated and not directly written in a composite measure lingo, is short and good value for the page.

Debates on composite measures versus dashboard presentations are ongoing in the international development community, particularly among World Bank econometricians; we owe an intellectual debt to some of them (Ferreira and Lugo 2012) for encouraging us to try out alternatives. Although this paper is about multidimensional poverty analysis, the methodologically interested reader will find it easy to see the relevance of their "middle ground" recommendations for the humanitarian needs assessment field.

Introduction

This note discusses composite measures in the context of impact and needs assessments.

A composite measure is a variable based on several constituent or original measures or variables combined into one outcome variable. In the process, the constituent variables undergo transformations (normalization, weighting and aggregation) that allow the computation of one number per measured item. With several properties condensed into one measure, the items can be compared in an ordered dimension.

Composite measures appear under a variety of broad quantitative construct terms and serve interests in and across many institutional fields. Economic indices (singular: *index*), social-science *scales*, or psychometric test *scores* exemplify predominant, but not exclusive terms by which they are known in the respective professions and professional audiences.

Interest in composite measures has surged in the last ten years, perhaps as yet another result of globalization, which makes numbers easier and cheaper to translate than text, and certainly much easier to aggregate, at the price, though, of more serious problems of validity and, ultimately, understanding among participants.

While some composite measures have been around for a long time, such as the global financial indices, some of the newer ones have suddenly reached top billing, chiefly because they upset cherished assumptions by presenting new evidence in a hard and apparently compelling way. For example, the Program for International Student Assessment (PISA), which administered its first tests in 2000, propelled Finland to widespread admiration, only to be outdone by Shanghai, China, in 2009.

The institutional variability of composite measures is almost endless and cannot be pursued much further here. Before we turn to their place and feasibility in rapid needs assessment, however, three general considerations from organizational sociology are worth stating here:

1. **Dominance:** Composite measures have been propagated by the revolution in accountability. There is hardly any institution left that is not under pressure to demonstrate performance by formal, quantitative measures. When these are ranked, they circulate easily ("Finland is the best!"), making remote surveillance possible. In other contexts, the discipline that they impose is milder. In economic policy, for example, the consumer price index is one of the bases of the inflation rate and is thus eagerly monitored. However, the index has to be evaluated together with a number of other key indicators. It cannot be used to abbreviate the health of the economy in one number. A major distinction, therefore, is between systems that are tightly coupled (e.g., US college ratings - media - prospective students) (Sauder and Espeland 2009) and those more loosely tied, in which no single measure dominates debate and decisions.

2. **Scientific basis:** A second point concerns the scientific underpinnings of composite measures. Composite measures like the Gross National Product, certain psychological personality inventories or the Human Development Index have been developed over a long time, with extensive attention to validity and reliability. They are integrated into the conceptual systems of particular sciences. As variables, they are routinely used alongside many others in models that test for the causal effects that the theories of the field posit. As data, they are comparable across populations and points in time because care was taken to standardize measurements.

   Distinctly from scientifically controlled fields, social movements may need composite measures in order to reduce information complexity, but they have to improvise them from available data without much knowledge of the underlying processes. The Global Landmine Survey, for example, an outgrowth of the movement to ban anti-personnel mines, classified mine-contaminated communities by a measure that combined ordnance, resource blockage and victim information. The "Mine Impact Score" employed a "weak metric" for combining some fifteen original measures. The weak metric was accepted because the behavior of affected communities was poorly understood, and only types, not the extent, of blocked resources could be measured under time pressure.

3. **Time pressure** is critical on its own. Composite measures may be established as part of slow-moving systems. Alternatively, they may have to be improvised under time pressure. PISA flowed from educational evaluations undertaken since the 1950s. As an organization, it had three years to prepare for its first assessment. By contrast, a traveler who realizes at the railway station that his credit card lets him withdraw only Euro 1000 for the next two weeks has ten minutes to figure a minimal budget - a composite measure of quantities (hotel nights, meals) and prices (hotel rates, etc.) - before deciding to board the train or to stay home. The OECD Handbook on Constructing Composite Indicators (Nardo, Saisana et al. 2008), to which we will refer extensively, sets standards for good composite measures. However, its didactic example - the Technology Achievement Index - makes it all too clear that it developed its guidance for researchers in dynamic, but not utterly rushed work environments.

We will situate the case for, and the evaluation of, composite measures for rapid needs assessments in the context of how much a particular measure should be allowed to dominate perceptions, whether there is a scientific basis for it, and how the gold standards of composite measurement design may be lowered under time pressure.

We will also try to develop an alternative to composite measures. Movements trigger counter-movements; the worldwide frenzy to impose metrics on each and everything has run into resistance. "Can a single performance metric do it?" (Kukla-Acevedo, Streams et al. 2012) formulates these doubts for the education sector, but may just as well be asked of any other sector falling under the spell of one-tells-all measures. Closer to our realm, a debate has grown louder in the international development community (Ravallion 2010).

Ferreira and Lugo (2012) sketch out a middle ground between indicator dashboards and what others contemptuously call "mashup indices". This latter article in particular encouraged us to demonstrate the use of different impact subindices. While reducing the information complexity, subindices preserve a useful degree of impact diversity.

Organization of this study

The remainder of this study proceeds as follows:

The next chapter discusses the various motivations for composite measures in rapid needs assessments, as well as the ambiguity about what they really express. We go on to present the OECD "Handbook on constructing composite indicators", which some may consider a kind of gold standard. Its chapter structure will guide much of our technical discussion. We situation the subject again in the humanitarian arena, by reviewing composite measures devised by the Global Focus Model project and by a rapid assessment of flood-stricken Cambodia, both of them UNOCHA-led endeavors. The Cambodia data let us demonstrate an alternative to the one-tells-all composite-measure approach. The practical guidelines are for assessment analysts who work on a spreadsheet platform; in small measure they also exemplify analysis forms required by more complex data and for which statistical applications are needed. An outlook section balances the pessimism of current practice with the optimism of future developments.

To lighten up the abstract and theoretical tenor of much of the general ruminations, the chapters on the UNOCHA projects and on the practical guidelines are punctuated with several sidebars. These stay close to the data, with graphs and also with formulas and practical hints that the reader may further investigate in the demo workbook. Two appendices elaborate on technical arcana of fine-tuning composite measures.

Composite measures in rapid needs assessments

Purpose

The context of our discussion here is impact and needs assessments. It is obvious that other elements of the disaster response may call for their own composite measures. A trivial example would be "tonnage of food and non-food relief" shipped to districts. Our concern is with measures relating to observations of the "X is affected by the disaster", "X is impacted in the sense of Y", "X has unmet needs in Y-area"-types.

A multitude of single indicators speaking to such conditions are widely used in humanitarian dashboards and other types of assessment documents. Assuming that dashboards permit quick, informative overviews, why would one want to combine the original indicators in composite measures? If only the composites are reported or consumed, obviously the assessment loses information.

There are several motives creating a demand for composite measures:

1. **Correlated impacts:** Disaster impacts are often correlated. The simple, parallel visualization of impact indicators in tables and graphs can highlight some of the

correlations, but with more affected entities and more types of impact the dashboard approach soon gets overwhelming.

2. **Ranking:** Until the various impacts are summarized in one measure, the affected entities are difficult to rank. Although it is often conspicuous which community, district or province is the by far most severely impacted, in the mid-field of affected units judgments on severity are more difficult, and a rankable composite measure appears to help reduce confusion and complexity. Entities ranked by some composite impact measure are much easier to represent in maps. They lend themselves to quick, if crude, geographical targeting of affected populations[1].

3. **Heterogeneous measures:** Composite measures may also permit the incorporation, in one impact expression, of measures that were handled differently by participants of the assessment process. Suppose that in one region assessors measured the sanitation impact by the percentage of communities with damaged sewer systems, in another region by the number of wells flooded with contaminated water. Some composite-measure formula may be devised to incorporate both indicators for a nationally comparable impact measure.

4. **Reliability:** Similarly, composite measures may be employed in hopes to minimize measurement error. The same underlying concept may be shed light on by including two or more indicators presumably closely associated with it. Thus food deficits might be measured by estimates of crop loss and additionally by changes in the price of staple food in nearby markets. Assessors may anticipate that both measures are unreliable and may hope to reduce error by combining them.

The legitimacy of the motives decreases from 1. to 4. The correlation of impacts *must* be addressed because multiple impacts have disproportionate effects on human suffering and on the chances of recovery. As for ranking, finely-grained differences on one summary measure may be less informative than a few broad groupings defined on one or two dominant (from policy interest) measures among the initial ones. Combining separate measures (where one is missing when the other is used, and vice versa) is legitimate if they speak to the same underlying reality and the relative weights are plausible. Finally, measurement error can be neutralized with the help of redundant measures only if their errors are independent of the true values[2], and are not correlated among themselves. This is hardly the case in major disasters since devastation and measurement troubles escalate together.

[1] Rankings may develop their own undesirable dynamic. As a reviewer put it: *"In the context of rapid needs assessment, there is a danger that the location initially considered the worst impacted (often on the basis of a single incomplete data set with limitations resulting from methodology) can remain the focus of both response and assessment activities well beyond the point where other data sources start to show other areas as severely, if not worst, impacted".*

[2] So-called classical measurement errors (Chen, Hong et al. 2011: 902).

Basic constellation

So far, we have not yet built a strong case for composite measures. At the least, there remains a need to keep the complexity of the initial information accessible and considered when forming such measures and again when drawing conclusions from them. This has been recognized by practitioners such as the food security community[3].

We now consider a basic situation that faces most rapid needs assessment, and from which we develop further questions and some conceptual notions to transition to actual composite measures used by others.

Two views - "goods" and "bads"

Figure 3 presents a notional view of the pre- and post-disaster situation affecting a country. For the sake of discussion, A, B and C stand for provinces. The (many) welfare dimensions are sampled here by two: habitat, measured per the percentage of families living in homes with a minimum surface per member, and education, for which some enrollment ratio is shorthand. Province A and B represent the development frontier; C exemplifies less-developed provinces.

A disaster occurs, setting all of them back in both dimensions, although in different measures and proportions.

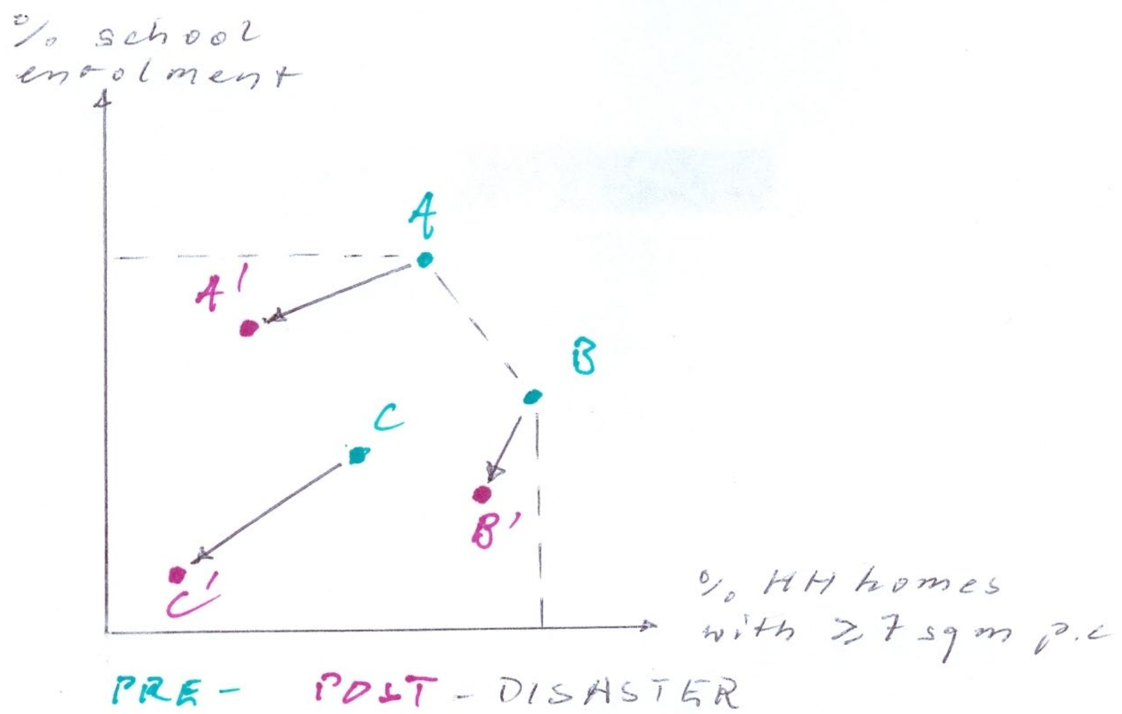
---

[3] Consider these statements from an Integrated Food Security Phase Classification manual (IPC Global Partners 2008: 21):

"Convergence of Evidence: Although the IPC strives for objectivity and consistency, the extremely complex nature of food security analysis makes the strict application of single indicator thresholds both impractical and technically questionable in their application to a wide array of situations. To overcome this, the IPC supports a Phase classification statement based on convergence of evidence from multiple sources (not limited to single assessment findings) as evaluated by analysts. Analysts use the reference outcomes as a guide, but ultimately make a classification statement based on the convergence of evidence from all available sources. This can include direct and/or indirect1 evidence of outcomes from a variety of sources and process indicators, depending on data availability and practicality.

Mixed Signals of Indicators: Given the complexity and diversity of food security and humanitarian situations, individual indicators may not consistently support the same Phase Classification. While this is a practical reality, the approach of the IPC is to make these differences explicit, examine them in their broader context and strive to make an overall Phase Classification statement using a convergence of evidence. Any notable deviations for particular indicators will be highlighted in the Analysis Templates, and should be explained."

Figure 3: Welfare GOODS view

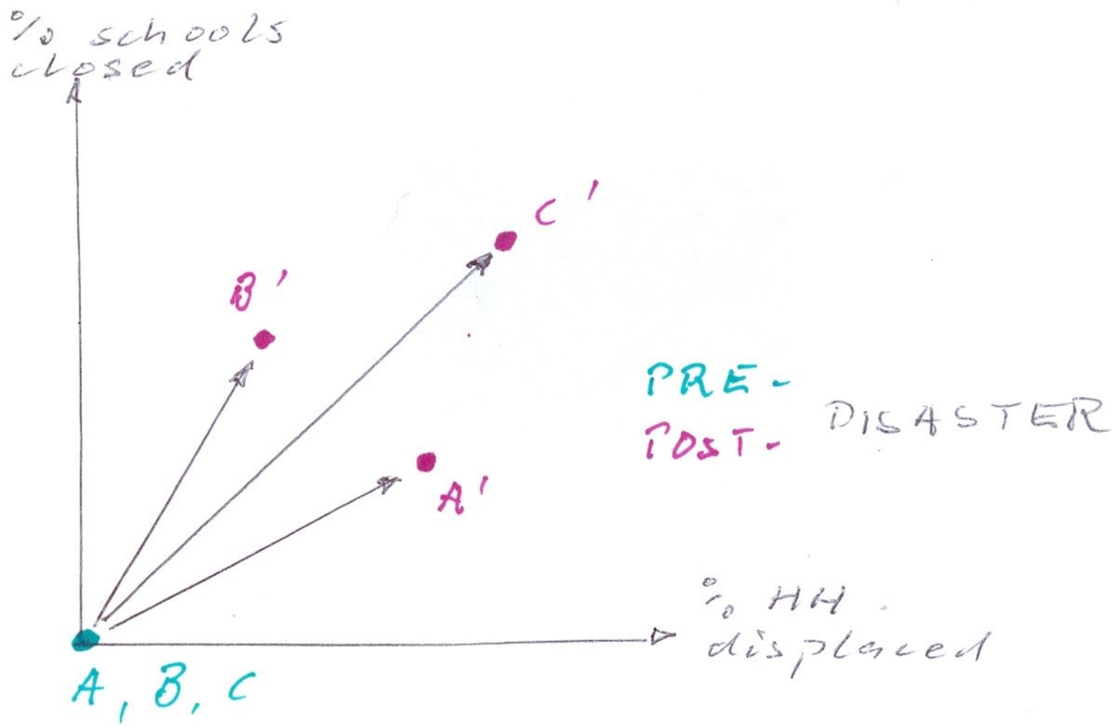


While the pre-disaster information may be known to the needs assessment team (and if so, it may be very valuable), the post-disaster values can generally not be rapidly measured in the same definitions. Even if that were theoretically possible, the same measures might not be directly relevant to the response planners.

Instead, typically, we switch from measuring welfare *goods* to welfare *bads* - variables that express loss, want, damage, pain, dislocation, etc[4]. In our construed scenario, the habitat situation may now be captured as households displaced from their normal homes, and the education situation as schools closed because of the disaster. However, displacement expresses more than just loss of habitat; it also causes disease, job loss, and other negatives. In the same logic, closed schools are not only an education event; they likely go hand in hand with other kinds of service disruptions. Thus, not only have our measures changed from "goods" to "bads", but they also cover different, though overlapping, institutional complexes from the pre-disaster measures. Figure 4 indicates changes on measures that are observable (and deemed relevant) by the assessors, assuming that all pre-disaster values of these bads were zero.

[4] For lack of better terms. *"It may be easier to reach consensus about welfare bads than about welfare goods"* (Offer 2000: 29). Closer to our subject, "undesirable output model" or "bad output model" (Cooper, Seiford et al. 2007: 367 sqq.).

Figure 4: Welfare BADS view



Coming back to the composite measures, it is obvious in this stylized situation that province C is the most severely impacted. Howsoever we weight displacement and closure of schools, C' will always exceed A' and B' on both dimensions. It is equally obvious that the relative ranking of A' and B' regarding their total impact (calculated as some composite measure) depends on the relative weights given displacement and school closures. This basic insight will follow us to the practical considerations later.

Complexities in the short-term view

Several important questions arise immediately:

1. **Impact typology:** Do we really want / need to rank A' and B'? (ranking C' versus the others seems unproblematic). What does this achieve for the needs assessment? Is it more helpful to identify constellations of impacts, such as "highly impacted in both dimensions" (exemplified by C'), "high population displacement, but mild service disruptions" (A'), etc.? (including the hypothetical D': "only mildly impacted on both dimensions", not shown here).
2. **Weighting:** If we decide to combine the impacts in one measure, the question of weighting arises (we omit questions of normalization and aggregation for the moment). Who sets the weights? What justifies particular weights? Should impacts that are not immediately life-threatening receive positive weights?
3. **Denominator:** In our example, the impacts are denominated to pre-disaster aggregates (proportion of households displaced now to households resident in province before; schools closed now to schools presumably functioning before).

How will the picture be different if we use absolute figures? Similarly, for geographical concentration within entities: The percentage of displaced households in the entire province may seem low, but if only a few districts are affected, it will be high locally.

[Sidebar:] Population-denominated indicators - yes or no?

Disasters hit some individuals and communities hard, others less so, yet others not at all. Impact concepts, therefore, are naturally understood as intensities. Operationally, these call for rates, ratios, or proportions.

Three types of measures are candidates to figure in such quantities:

- The (unrated) impact (e.g., casualties) (I)
- The affected population (e.g. those living in flood-affected villages) (A)
- The exposed population (e.g. those living in the provinces with a history of flooding) (E)

Two challenges are obvious:

Impact, participation and exposure all depend on thresholds. In the example of casualties, this is the difference between "dead, injured or missing" versus "unharmed and accounted for".

Second, the data available (or collectable within useful time and effort) express them imperfectly. Data limitations then determine which of the three potentially useful ratios can be computed:

- I / A
- I / E, or
- A / E.

In the needs assessment perspective, to the extent that they can be measured in fairly complete and undisputed manner, I / A seems preferable. It holds a measure of loss, suffering or unmet needs (the numerator) to a measure of those who are the sufferers (the denominator). If the cases can be broken down on some other dimension (groups, areas, sectors), profiles of impact intensity emerge.

Other professional perspectives - such as disaster risk mitigation - may privilege I / E, in an effort to learn from disasters as they materialize amid populations defined by risk types. Finally, A / E may be of greater importance in development, where A may stand, for example, for the difference in headcounts of poor people post-/pre-disaster (those who fell into poverty regardless what the specific disaster impacts are) and E for all people living in a chronic disaster area.

For the purpose of composite measures, let us assume that I / A measures are feasible. Then three considerations apply:

1. All indicators should be ratios. Mixing ratio indicators with unrated ones causes bad validity issues.
2. Depending on the nature of the impacts (the numerators), the denominators may vary (e.g., households for displacement; total farmland for silted surfaces).
3. The uncertainty of the ratio measure is a compound of the uncertainty in numerator and denominator. Only if both measurements are biased in the same direction, it decreases (compared to uncorrelated errors).

The first point is restrictive; the second may open more possibilities; the third suggests that the composite measure designer should think hard about plausible measurement errors. In practice, key informants may, for some types of impacts at least, volunteer ratio-type estimates anyway, such as "About half of all people have moved to shelters", "Almost three quarters of our land cannot be planted in the next season."

From an impact measurement perspective, population-denominated indicators are attractive. Their feasibility depends on circumstances. They are opportunistic measures for which we should look out, but which cannot be mandated.

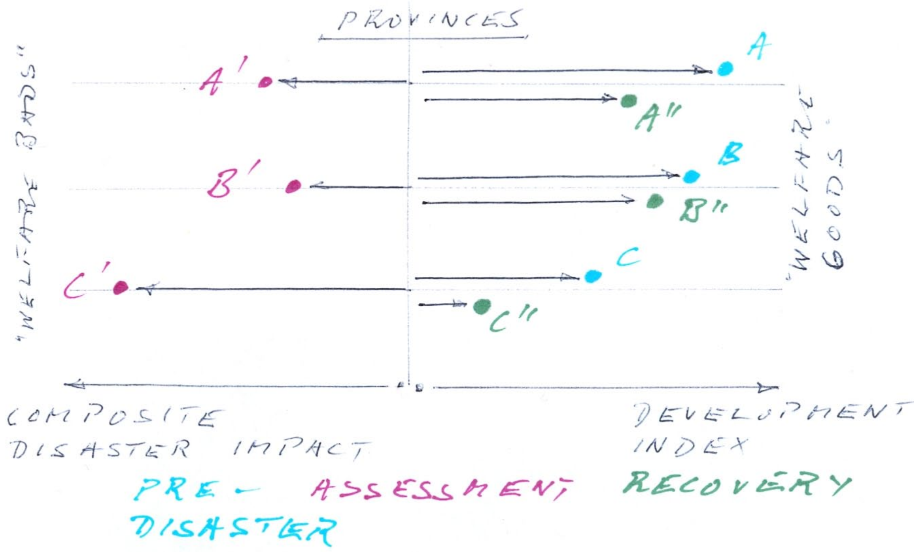
Longer-term complexities

Those are validity concerns that face designers of composite measures in the short run. There are others that come to the fore when we take a longer perspective. While welfare goods are not observable immediately after the disaster, we would nevertheless expect that a composite measure formed of "bads" will be meaningfully associated with the speed and extent of recovery. Recovery ultimately will be measured in terms of welfare goods. Ideally, with full information, recovery levels would be predicted as a function of pre-disaster development, disaster impact, and relief and rehabilitation activities. Even with incomplete information, a relationship between impact and recovery measures should be traceable. In other words, we expect the impact measure to have some predictive validity vis-à-vis outcomes that occur much later[5].

This figure suggests such a relationship when we compare provinces A and B to C. C is the one suffering the biggest disaster impact as well as showing the largest recovery gap. The fact that C was also the least developed before the disaster likely contributed to this outcome. With data on a sufficient number of affected units, recovery levels could be regressed on pre-disaster conditions as well as on some or all of the base indicators in the composite impact measure. The indicator coefficients could then be considered criterion-validated weights. These weights could be re-used in later assessments of similar disaster. In fact, such analyses might be attempted already with data collected under Phase-4 of common needs assessment, as part of the "periodic multi-cluster monitoring" (Garfield, Blake et al. 2011: 4).

[5] Hurricane "Katrina", which devastated Louisiana and particularly the city of New Orleans in the southern USA in 2005, spawned some research along similar lines. For example, Finch, Emrich and Cutter (2010) measure recovery by the repopulation level three years later. Their disaster impact measure, however, is a single variable (though combining several measurements in the same neighborhoods) - the depth of flood waters measured remotely on the day of near-maximum flooding. The pre-disaster measure is a composite measure of social vulnerability.
This is basically a model connecting three "bads" - social vulnerability, physical devastation, and (the complement of) depopulation. With data on 181 census tracts in New Orleans, the main finding was that the correlation between impact and recovery was strong (-0.69), but that between vulnerability and recovery only modest (-0.25). Vulnerability and impact were virtually uncorrelated. The vulnerability effect on recovery is in part masked by the distribution of government aid, which was targeted to the poorest areas, delaying returns of middle-class households. The study is particularly innovative for the use of a proxy measure for census tract repopulation (in the absence of a proper census in 2008) - the volume of postal deliveries apportioned from delivery routes to the census tracts!

Figure 5: Development, disaster, recovery - composite measures



In reality, this is unlikely to work, not least because the response to subsequent disasters learns from the earlier ones (Comfort, Oh et al. 2009). We are left with the question what it is that composite impact measures in rapid needs assessment are meant to express:

- The short-term needs of individuals, or
- Stress on short-term societal functioning, or
- Anticipated longer-term recovery needs

"Needs assessments" lean towards the first understanding, with recognition also of the needs of essential institutions, not only individuals. The major challenge, however, is that we do not have a ready mechanism to decide the importance of the different indicators in short-run perspective whereas a longer-term retrospective would allow some kind of validation for different weights. This challenge will not go away; there will always be doubts whether "a single metric can do it all". The doubts motivate the search for alternatives to one composite measure.

[Sidebar:] Indicators as one of several data types in assessments

So far, we have implied that the base indicators of composite measures are interval-level continuous or count variables. This seems desirable, at least for consistency. But such indicators are not the only types of variables that are generated in rapid needs assessments, and may not be the most important or the easiest to measure.

A typological approach to impact measures may distinguish them by the way they express needs (absolute vs. comparative), and by single vs. multi-sector content. In this scheme, impact indicators in their majority fall into the absolute need-cum-single sector combination. Admittedly, some indicators (e.g., displacement) speak to several sectors. And, once they are combined, and a ranking is establish on the basis of a composite impact measure, we use them comparatively.

Table 1: Typology of impact data

TYPE OF MEASURES	Salience of need (absolute)	Priority of need (comparative)
Across sectors	*TYPE 1: Scale items (dichotomous)*	*TYPE 2: Ranked priority sectors (ordinal)*
Within sector	*TYPE 3: Impact indicators (interval-level)*	*TYPE 4: Severity rated key issues (ordinal)*

Examples of the other types of data are:

- Type 1: An adapted version of the so-called Hesper Scale (WHO and King's College London 2011), with problem items (e.g., "Is water supply a major problem in this community?") used to position communities on a common scale.
- Type 2: A true priority ranking of sectors by community key informants.
- Type 4: A severity-based rating of problem, elicited by sector, formulated and evaluated by communities.

The combinations of measures of different types hold a potential that has barely been investigated. Type-3 indicators are "objective" measures, such as lives lost, that can be rated to meaningful denominators. From their correlations the analyst can identify diverse needs profiles. These then need validation against the affected communities' own preferences, as expressed through constructs of the other types.

We now leave these bare-bone conceptual musings in order to reference a scholarly work that looms as a kind of gold standard for composite measure design in the relevant literature. Thence we will move on to applications in the humanitarian area.

The OECD handbook: The orthodox position

The "Handbook on constructing composite indicators" (Nardo, Saisana et al. 2008), published by the Organization for Economic Co-operation and Development (OECD), is a detailed, much-cited[6] resource for this purpose. Although it does not claim any such status, many users may be regarding its prescriptions as a kind of gold standard. However, the handbook is written for research, policy analysis and advocacy contexts that differ in important ways from rapid needs assessments post-disaster. It presumes a data-rich, highly analytical, slow-moving research environment. Many of its prescriptions, particularly of the analytic type, are not feasible in the typical situation of interest here.

[6] Google Scholar, as of 30 March 2012, claims 320 citations for its 2005 edition, and another 46 for the 2008 edition.

Yet, the handbook is valuable as a roadmap that lays out steps in construction and validation of composite measure. It is useful also because it raises a number of warnings that assessments teams ought to know even if, again, not all of them may be practical.

The handbook cuts the design process into ten steps (pp.15-16), each of which is awarded its own chapter:

1. Theoretical framework
2. Data selection
3. Imputation of missing data
4. Multivariate analysis
5. Normalization
6. Weighting and aggregation
7. Robustness and sensitivity
8. Back to the real data
9. Links to other variables
10. Presentation and visualization

We will revert to these in the practical guidance section. Here we will briefly list some of the points that it presses:

Transparency: Having a theoretical model of the subject to which the composite measure speaks is important. This may be best developed in economics, where an endless variety of composite measures rely on quantities multiplied by natural weights (their market prices). In areas, however, that are new and poorly understood, the theoretical underpinnings will be weak. This is the case also in rapid assessment situations, where 1. the behavior of the distressed society is predictable to a minor degree only, 2. good data are scarce and therefore drive, rather than being driven by, indicator design. In such situations, transparency is a critical virtue so that peers and consumers understand what was put together and how it was computed, and may possibly improve on it if better data becomes available in time[7].

Weights: Like most other critics of composite measures, the Handbook objects to the arbitrary nature of the weighting process. It sees this as a major challenge to validity, particularly in so-called compensatory methods. These apply to designs in which the effect of the reduction in one contributing variable can be offset by an increase in others. Methods that preclude such trade-offs are called non-compensatory; they are clearly favored by the authors of the Handbook. However, these methods are analytically difficult, and the one highlighted in the Handbook has not been implemented, to our

[7] If the call for transparency sounds like a platitude, consider again the case of the Gross National Product. We do not need to know how it is measured in order to take a serious interest in the growth rates on which it is based. We trust that enough competent economists have seen to its valid and reliable measurement.

knowledge, in any of the applications commonly used by assessment teams[8]. This leaves us with compensatory methods as the practical "second best".

Advocacy vs. response planning: In their final remarks, the Handbook authors express their opinion (p. 137) that *"an aggregate index* [i.e., a composite measure. AB] *might be useful to make an argument for action"* while *"for policy formulation"* [in our situation: needs assessments and response planning] *"individual variables and quantitative analyses .. might be more relevant"*. The aggregate index permits ranking, such as of areas more or less severely impacted, and thus can be used to mobilize stakeholders and publics. Analyses of the non-aggregated individual variables permit to drill down and to precisely determine interactions. They are thus more useful for differentiated action planning. With these second thoughts, the Handbook adds its own bit to the dashboard vs. indices debate.

Examples from the humanitarian area

As a transition to the guideline section, we present two composite measures used, respectively created, in the UNOCHA network. These are: 1. the country risk score calculated in the 2012 version of the "Global Focus Model (GFM)" (UNOCHA and Maplecroft 2011) and 2. a needs assessment in Cambodia after the floods of summer 2011 (UNOCHA 2011) .

Only the second is a direct outcome of a rapid needs assessment. However, both illuminate a number of interesting points that matter in our context. If we were to compare them to some of the more established and household-word composite measures, e.g. the Gross National Product, some salient differences would leap to the eye.

Table 2: Three composite measures compared

Institutional sector / composite measure	Theoretical basis	Analytical units	Weighting	Speed of execution
National statistical offices: GNP	Economics	Firms and consumers	Natural (market prices)	Usually quarterly updates
Humanitarian: GFM country risk	Risk = f(hazard, vulnerability, capacity)	147 countries	Arbitrary (multi-level indices)	Annual
Humanitarian: Cambodia post-floods	Data-driven	16 provinces	Arbitrary (sectors get equal weights)	2-3 months

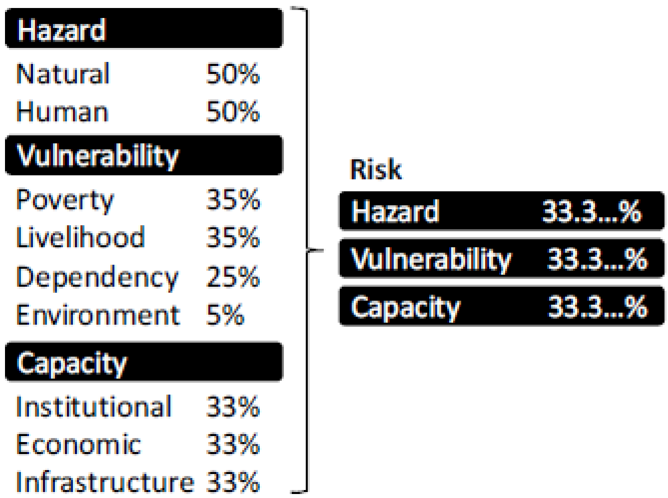
The Global Focus Model

An UNOCHA venture, the Global Focus Model produces annually updated risk estimates regarding humanitarian emergencies that might befall the 147 countries assessed. The

[8] The Condorcet-Kemeny-Young-Levenglick (C-K-Y-L) ranking procedure (p.110 and passim). An Excel add-in performing it does not seem to exist. For an older, yet brief and accessible introduction to non-compensatory methods, see Yoon and Hwang (1995).

risk score is a composite of three variables given equal weight - hazard, vulnerability and capacity -, which in turn are composite measures at an intermediate level. Figure 6 (UNOCHA and Maplecroft 2011: 5) shows the relationship at the three highest levels[9].

Figure 6: The Global Focus Model risk score composition



However, we are interested in the mechanics of the risk score as computed from the nine variables shown in the diagram. These again are composites each of several variables. The poverty index, for example, incorporates country population, night-time luminosity, national poverty and infant mortality data. We are not concerned with this lowest level of base indicators.

The indices were calculated such that their values fall between 0 and 10. The score for all variables is defined with 10 denoting the highest possible negative outcome, 0 the best. The empirical range runs from 1.6 for Singapore (most stable) to 7.7 for DR Congo (most risky).

The results are displayed in League Table format, with countries sorted descendingly (by Focus score) and colored by quartiles[10]. The risk quartiles are also displayed in a worldwide map, which makes the clustering of high-risk countries in sub-saharan Africa and South Asia conspicuous. A scatterplot of capacity scores by vulnerability scores hint at correlational analysis, which otherwise is not evident.

The report is compelling for its clear presentation, including of the hierarchical structure of the indices and of the weighting scheme. The theory underpinning the selection of variables (and presumably also their transformations) is borrowed from the well-known International Strategy for Disaster Reduction formula, whereby

$$\text{Risk} = \frac{(\text{Hazard} + \text{Vulnerability})}{\text{Capacity}}$$

The functional form, however, is not to be taken literally; the aggregation is by linear combination, i.e. purely additive. There is no division by capacity, but rather its subtraction, by the fact that the highest capacity is scaled to 0, the lowest to 10. Division would have made the calibration of the resulting risk score challenging.

[9] The risk score is not the ultimate composite that the model produces - this is the Focus score, which is the risk score adjusted by a humanitarian factor. We are not concerned with the Focus score.
[10] Actually, what Figure 3 on page 2 calls "quartiles" are four equal intervals of the theoretical range, rather than statistical quartiles in the usual sense of equal-frequency quartiles.

Two major questions arise that are of general interest for composite measures, and which will challenge the design and interpretation also of those used in rapid needs assessments:

1. How are the functional form and the weights chosen?
2. Do the participating indicators stand for processes that are distinct in reality, or are the differences mostly semantic? If there is significant redundancy between indicators, does the weighting system correct for this, or are the results biased to units that happen to be on the extremes of several redundant indicators?

The functional form - the purely linear combinations of indicators in this case - is probably determined by the rudimentary state of behavioral theory about emergencies and emergency response. With so little known about risk - except that it is plausibly fashioned by hazard, vulnerability and capacity -, the simple addition of scaled and weighted indicators is the most straightforward and easiest-to-understand form. One could, of course, have added other credible theoretical elements, such as the cumulative effects of multiple disadvantages. These may justify non-linear forms of aggregation.

We conducted such a simulation. We recalculated the risk, using a formula that rewards similar levels, and penalizes differences, among hazard, vulnerability and capacity indices, given the originally computed risk. In other words, in the old formula two countries with values (5, 5, 5), respectively (0, 5, 10) have the same risk score in the Global Focus Model, i.e. 15/3 = 5. The alternative formula, built to reflect cumulative disadvantage, raises the value of the first country relative to that of the second.

The result is instructive. The rankings of counties are very similar under both aggregation formulas. The mean absolute change in ranks between the two formula results is a mere 3.2 - little for a set of 147 countries! This suggests that, overall, the GFM risk is robust to alternative aggregations. The rank changes are most pronounced in-between the extremes, in the midfield. The most risky as well as the least risky countries score similarly on all three indices and therefore see their risk barely changed by the alternative formula. Curiously, this formula effects the largest rank changes among a small group of island nations, characterized by low hazard, medium vulnerability and fairly good capacity. The Marshall Islands, for example, with values of 0.5, 4.3 and 6.7, originally were awarded a risk score of 3.9, which drops to 2.9 under the alternative. Thus re-scored, the Islands' rank changes by 18 points.

We retain three tentative insights from this:

1. **Simple methods:** The overall rank structure of composite measures may be fairly robust to the choice of weights and aggregation functions (subject to verification in the specific situation!). This working assumption is in line with the results from multi-method comparisons in "Multiple Attribute Decision Making" (MADC), a field that has a lot in common with composite measures (Yoon and Hwang 1995: 68-69). If correct, then, obviously, this is a strong recommendation to use the most simple among all adequate methods for calculating a composite measure in a given situation. Added analytical sophistication is not worth the loss of time and

common understanding that it likely entails under the time pressures of rapid needs assessments.

2. **Robust extremes:** The composites are more robust at the extremes of the composite measure; they tend to yield to alternative specifications the most in the middle range. This finding is confirmed by robustness simulations that the OECD Handbook reports from an entirely different realm (op.cit.: 118, Figure 18). The reason is trivial and is found in the commutability of the additive operation: 0 + 0 (at the low end) is always 0, regardless of what weights are attached to the inputs. At the upper end, 1 and 1 give the same result whatever the weights. In the middle, however, f(1, 0) = f(0, 1) only for equal weights. This is both reassuring and limiting: the extremes will be detected reliably despite the uncertainty of the methods; the real differences among disaster-affected units may call for different policies, but composite impact measure may obscure them.

3. **The indicators matter:** As Yoon and Hwang (ibd.) emphasize, the most crucial part is in the "generation of appropriate attributes". In our language, this means that the decision about which indicators to use (and, if they are not available, on which to concentrate data collection) is more important than subsequent method choices.

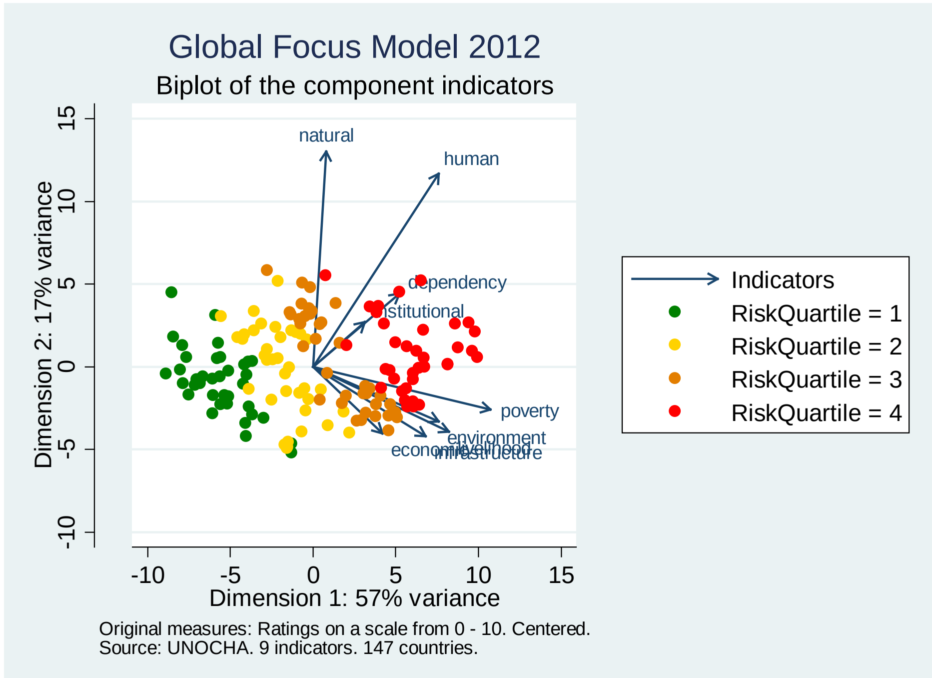
We look at this third point more closely through the structure of the indicators that the Global Focus Model used.

[Sidebar:] The internal structure of the Global Focus data

Composite measures result from indicator data that can be ordered in tables (matrices) of n rows (for n sample units) and m columns (m variables). Thus, the GFM data of concern here comes in nine variables (from "natural hazard" to "infrastructure" as shown in the diagram further above). It does not matter here that these nine are themselves composites of other, preceding measures.

Biplots (Wikipedia 2009) are a rapid and highly graphic method to represent relationships among the cases as well as among the variables. The biplot places information on both in a common coordinate system, with a maximum compressed to the two-dimensional plane. Points in close neighborhood indicate similar cases. Arrows forming small angles indicate highly positively correlated variables. Arrows perpendicular to each other indicated low or zero correlation. The meaning of the length of arrows varies with different flavors of biplots; in our case longer arrows mean that the two-dimensional representation is able to use more of the information in the variable.

Figure 7: Biplot of the nine Global Focus indicators



The biplot tells us two things of importance:

Variables: They roughly form two groups (or, as statisticians would say: factors). Group one, symbolized by arrows pointing to the right lower quadrant, consists of:

- Poverty
- Lack of livelihoods
- Environmental stress
- Lack of economic capacity
- Lack of infrastructure

The second group, pointing to the upper right quadrant, embraces the remaining four:

- Natural hazards
- Human-induced hazards
- Dependency
- Weak institutions

The two hazard variables are thus associated with the same factor. The vulnerability and capacity indicators, however, are distributed to both factors. So far, so good.

The question of general interest here concerns the possibility of substantive redundancy. In the case of the GFM indicators, one may ask whether the poverty and livelihoods variables indeed express different things, or more or less the same thing twice (their correlation coefficient is +0.87). This question is not about the original base indicators of which each is composed - it is about the underlying common concept (vulnerability in this case). Similarly, the economic capacity and infrastructure variables may be highly redundant vis-à-vis the underlying societal capacity concept.

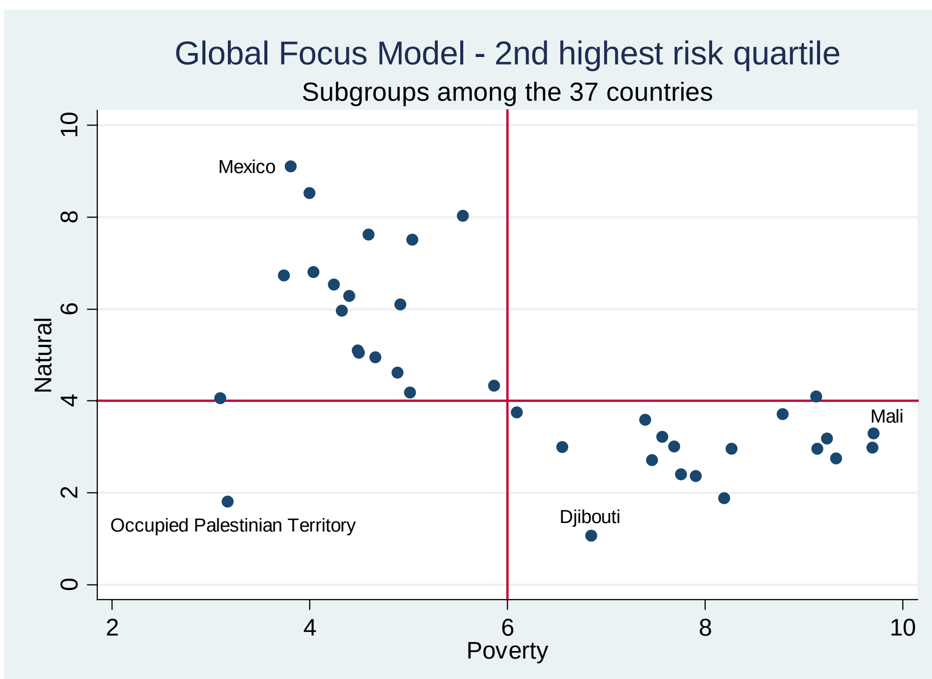
For the construction of the GFM risk index, this is irrelevant. The redundancy is almost neutralized by the weighting system. In other words, at present institutional, economic, and infrastructure capacity each contribute with a weight of 0.33 to the societal capacity index. If, to reduce redundancy, one or the other of economic or infrastructure capacity were removed, the remaining two would each be made to contribute with a weight of 0.5. The overall consequences would be minimal.

If these precautions via the weighting system are not taken, substantive redundancy may bias the results. The choice of indicators should carefully consider the presence or not of substantive redundancy. If more substantively similar indicators are available at little extra cost, they should be used. The weighting system then needs to reflect their number, in order not to give them overdue influence at the expense of other substantive areas. The GFM weighting system correctly takes care of this.

Cases: This biplot powerfully demonstrates that sample units that score in the middle range of the composite measure may belong to substantively distinct subgroups. The cases were colored according to the quartiles of the risk index. Of all the four groups, the members of quartile 3 (dark orange, second highest risk group) are separated into two distinct groups. Those below the y = 0 line are more closely aligned with the arrows of the first factor. Those above cluster around the natural hazard arrow.

To visualize this in more common-sense terms, we plot the values of the poverty and natural hazard scores for the 37 countries in this risk index quartile. The separation of the two sub-groups is patent, with only one clearly non-conforming member.

Figure 8: Substantively distinct subgroups in the middle risk range



Within the next lower risk quartile (yellow), there is an equal differentiation by poverty and natural hazards, but it is less distinct.

We summarize: The GFM structure supports the need for the designers of composite measures to

- consider substantive redundancy among the indicators and to control it either through the weighting system or by excluding redundant indicators

- identify substantively distinct subgroups among sample members scoring in similar ranges of the composite measure, particularly in the middle ranges.

The Cambodia floods of 2011

In 2011, the monsoon rains caused wide-spread flooding in Cambodia. Various government ministries, UN agencies and NGOs undertook needs assessments in the affected provinces between late September and the end of November. UNOCHA workers compiled this information, later publishing a report (UNOCHA 2011) that mapped out the physical extent of flooding, the areas visited by assessment teams, damage to the rice crop, affected wells, health centers and schools.

Indicators used

The UNOCHA team compared the flood impact across provinces by a composite measure formed from these four indicators:

- Displaced households
- Destroyed rice crop
- Affected health centers
- School needs

The school needs measure itself was a composite measure incorporating information on school pack, furniture and textbook needs. The number of affected schools was reported, but not included in this measure (we will use it in developing an alternative approach below).

Of the 23 provinces and the city of Phnom Penh (henceforward also considered a province), 18 were considered affected. Complete values on the four indicators were available for nine, and incomplete values for seven provinces. On the remaining two provinces, virtually all information was missing.

For most of our demonstrations using this data, we follow the UNOCHA team's decision to treat as zero all the missing values in the nine that returned incomplete information. This is a de-facto, if undeclared imputation. In a sidebar further below, we compare the original ranking of provinces to that obtained when other imputation methods are used.

Here we are concerned only with the mechanics of the composite impact measure, not with the other parts of the report. The report presents the full indicator data, including their normalized values (page 21), but the ultimate impact measure - the composite measure of interest - is barely used. It appears only once (and misleadingly is called "rank"), in a table in the section on geographical information gaps (page 26).

Simple transformations

Nevertheless, the design of the composite has a lot to recommend it for further study. The normalization, weighting and aggregation of the base indicators are clear and simple. Even though not much argumentative use of the impact measure is made in the report itself, it lends itself to illuminating some important aspects of composite measure design.

Normalization: The base indicators were transformed by dividing them by the country-wide sum of their initial values. This yields percentages by province of all the displaced households, damaged rice surface, affected health centers, and school needs.

Weighting: Each of the four transformed base indicators was weighted identically with the factor 0.25 (to ensure the composite measure too would add up to 100%).

The **aggregation** is by simple addition of the weighted transformed base indicators. In fact, *in this case*, weighting and aggregation are commutative; it does not matter whether we first divide each transformed indicator by 4, and then add them; or first add the indicators and then divide the result by 4.

Critical points

There are three observations to be made on the initial transformations:

1. **Sample-dependent normalization:** First, dividing by the sum ensures that even the smallest value greater than zero comes out with a positive normalized value. Division by sum is more appropriate to mapping disaster impact than, say, min-max normalization, which assigns zero to the smallest value. But it can be problematic in as much as the transformed values come to depend on the entire distribution. To see how this happens, consider a simple case of only three units and one indicator. In both scenarios, the extreme cases A and C have the same original values. B is different. Since B too influences the normalization denominator (the sum of the original values), as a result A = 0 remains unaffected, but the normalized value for C changes significantly.

Table 3: Stylized example for distribution-dependent normalization

Province	Distribution 1		Distribution 2	
	Orig.	Normal.	Orig.	Normal.
A	0	0.00	0	0.00
B	2	0.17	8	0.44
C	10	0.83	10	0.56
Sum	12	1.00	18	1.00

This, by itself is tolerable - all normalization schemes have pros and cons -, but it implies that there is no natural basis for equal weights. For setting weights, inspecting histograms and comparing medians (the means are identical, by construction) of the normalized indicators will be helpful - besides the substantive considerations.

There are alternative normalization formulas, such as dividing the original values by their maximum. We will list more of them in the guideline chapter.

2. **Population:** One might consider standardizing the indicators to provincial population size. This is intuitively attractive if we agree that 10,000 displaced persons in a population of 50,000 is not same as 10,000 amid 500,000. These ratios would then still have to be normalized in a second step, but they would take account of population size differences. This has not been discussed in the UNOCHA report, and in fact it would be a poor choice in the absence of more finely-grained information on districts within provinces. A province may be affected by the disaster only in a fraction of its territory. Using the entire provincial population as a first normalization could thus be greatly misleading. But, in other disaster and information contexts, it will always be worth asking whether population-standardized indicators have value (see page 17).

3. **Unit weights:** The UNOCHA analysts used equal weights for all four normalized indicators. This may facilitate consensus - no sector can complain about unreasonable discrimination. Yet, is everything equally important in a disaster situation? Should, in our case, school needs be given the same importance as the displacement of families? Rather, one would think that indicator weights should follow substantive reasoning.

Robustness to weight changes

However, we cannot engage in discussing the weights question *substantively*. Instead, on purely *formal* grounds, we show how the ranking of provinces by impact scores is affected when the weights given the base indicators change. For demonstration, we vary the weight of the school needs measure over the interval [0, 4]. Such a simulation table can be produced rapidly in Excel; details are given in the appendix and in a demo workbook.

Table 4: Robustness of rankings on composite measure to varying indicator weight

Province	Weight of the education needs indicator								
	0	0.5	1	1.5	2	2.5	3	3.5	4
Prey Veng	16	16	16	16	16	15	15	15	15
Kampong Cham	15	15	15	15	15	16	16	16	16
Kampong Chhnang	14	14	14	14	14	14	13	13	13
Kandal	11	11	13	13	13	13	14	14	14
Kampong Thom	12	12	12	12	12	12	12	12	12
Banteay Meanchey	13	13	11	11	11	11	10	9	8
Svay Rieng	10	10	10	7	7	7	7	7	7
Pursat	4	7	9	10	10	10	11	11	11
Battambang	6	8	8	9	9	9	9	10	10
Siem Riep	9	9	7	6	6	6	6	6	6
Takeo	7	6	6	8	8	8	8	8	9
Phnom Penh	8	5	5	4	4	4	4	4	4
Kratie	5	4	4	5	5	5	5	5	5
Kampot	2	3	3	3	3	3	3	3	3
Preah Vihear	3	2	2	2	2	2	2	2	2
Stung Treng	1	1	1	1	1	1	1	1	1

Note: Here, as elsewhere in our analyses of the Global Focus and Cambodia flood data, rank 1 stands for "least affected".

Two things leap to the eye:

1. Ranks change significantly only for some of the provinces that originally were in the midfield. At the extremes, the impacts of the disasters are more highly correlated; changes in weights will affect their composite measures (slightly), but barely the relative position vis-à-vis other provinces.

2. There can be rank reversals. Here we notice them for Takeo and Kratie, with, e.g. Takeo going from rank 7 downward to 6 and then upward to 8. Reversals are difficult to explain. Ranks are not statistically independent between units. Moreover, when we consider the composite measure as a function of the weight of a particular indicator, given indicator values and proportionate adjustments in the weights of all other indicators, then the measure is not linear in the weight. There is no need to elaborate further. The point is that rank reversals in such simulation tables are no reason for alarm.

Good qualities

We summarize the insights gained so far from the composite measure:

- **Simple:** The initial normalization was simple - just dividing by the column sum - and preferable to a min-max normalization, which sets the minimum to zero. This ensures that, if weights remain equal for all indicators, all indicators are equally important in the resulting composite. "Equal importance", however, is a substantive decision. It implies a kind of "Everybody has won, and all (sectors) must have prizes"-philosophy, as in Alice in Wonderland. Designers of composite measures should at least make this explicit.

- **Commutative:** The aggregation was equally simple, by addition and division by the sum of indicator weights (4 * 1 = 4). In this linear set-up, weighting and aggregation are commutative, i.e. it does not matter whether the division by 4 occurs at the weighting or the aggregation stage. Note this is not the case in non-linear aggregations.

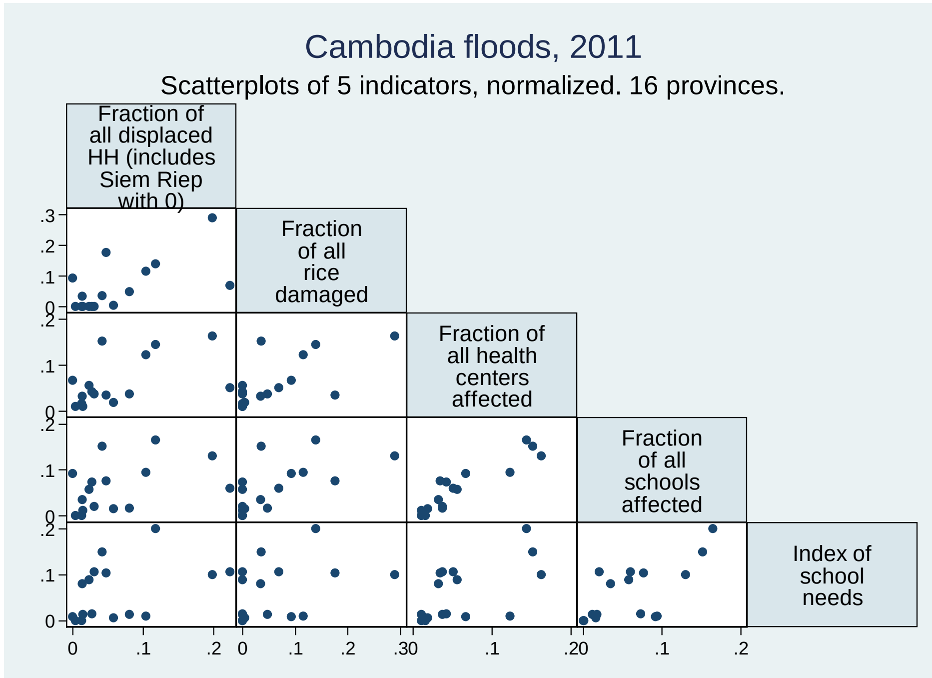
- **Robust:** The robustness to weighting was exemplified by varying the education needs weight. The ranking changes significantly only in the midfield of affected provinces. Even then, the changes appear modest - only Pursat Province changed from a low rank to medium-high. In other words, as far as we tested, the composite measure is fairly robust to weight changes.

At this point, we feel, the useful discussion of the Cambodia flood assessment reaches its limit, as far as the composite measure goes. We will, however, investigate an alternative to composite measures with this data, after a sidebar that visualizes the correlation structure of the indicators.

[Sidebar:] Network representation of indicator correlations

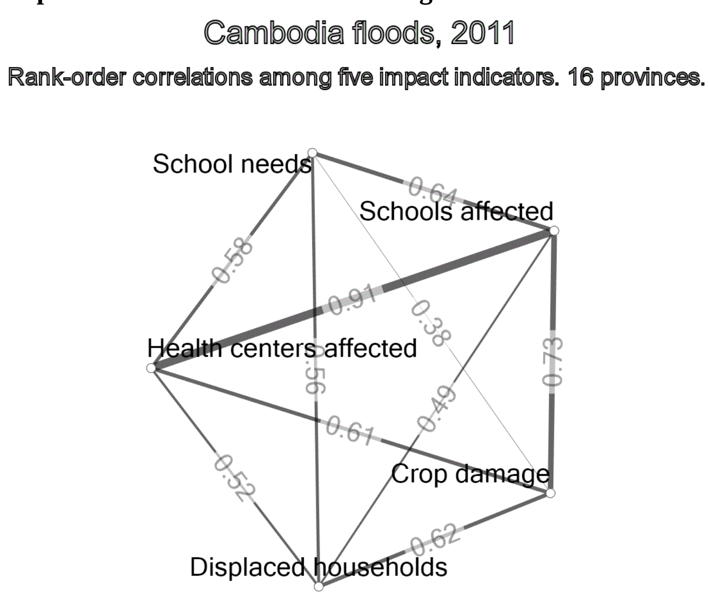
Scatterplots of the indicators developed in Cambodia show associations of greatly varying strength. Also, the observations are concentrated in the lower left corners of the panels. For this demonstration, we included also the number of school affected (which is not among the original four indicators combined in the composite).

Figure 9: Matrix graph of five indicators in the Cambodia floods assessment



Ordinary Pearson correlations will be dominated by the leverage that the few highly affected provinces exercise in the total picture. To do justice to the peculiar distributions, rank-order correlations may be more appropriate. We present them graphically in network shape. The thickness of the edges increases with the correlation coefficient.

Figure 10: Network representation of correlations among indicators



A very strong association can be noted between health center and school impacts. This may seem obvious, in the sense that the floods likely disrupted the operation of these facilities to similar degrees. But it may just as easily be the result of shared reporting mechanisms, whereby estimates of affected health centers and schools were elicited from the same local authorities. If so, we should assume highly correlated measurement error in these two indicators.

Conversely, the correlation between crop damage and household displacement is considerable, yet lower than one might expect. Inspecting the matrix graph, we notice that the province with the highest displacement reported relatively low crop damage. One might assume that farming households generally move when floods are so high that their crops too are badly damaged. If so, we should expect the two indicators to be strongly correlated. The observed correlation is modest either because our assumption is wrong, or because of different degrees of urbanization, or because of separate reporting channels.

We should therefore design composite measures in an awareness of how the indicators were measured, and whether strong correlations plausibly reflect reality, or are the result chiefly of correlated errors. Pervasive measurement error should be the default assumption and should not carry the stigma of poor work or other moral connotation. They are normal in disasters.

Excel's function CORREL computes Pearson correlation coefficients. The Data Analysis tool (in the menu: Data - Analysis - Data Analysis - Correlation) produces static Pearson correlation matrices. Calculating rank-order coefficients in Excel is tedious. The network graph was produced with the Excel add-in NodeXL (NodeXL Team 2012), which can import matrices.

A substantively motivated alternative

Here we briefly develop a substantively motivated analysis of the Cambodia flood indicators, as opposed to the equal weights-driven approach that the UNOCHA team followed. This may inspire similar treatments of other assessment data bodies.

Differentiated impact spheres

We start from the consideration that most disasters are unitary events. A bridge collapses; a catastrophic monsoon inundates Pakistan; drought pushes Sahelian populations into severe food crisis. The limits of this assumption are obvious. The Japanese tsunami and nuclear disasters, while causally and geographically connected, triggered distinct courses of events and response. The Tuareg rebellion in Mali is fueled by regime change in Libya as much as by the drought.

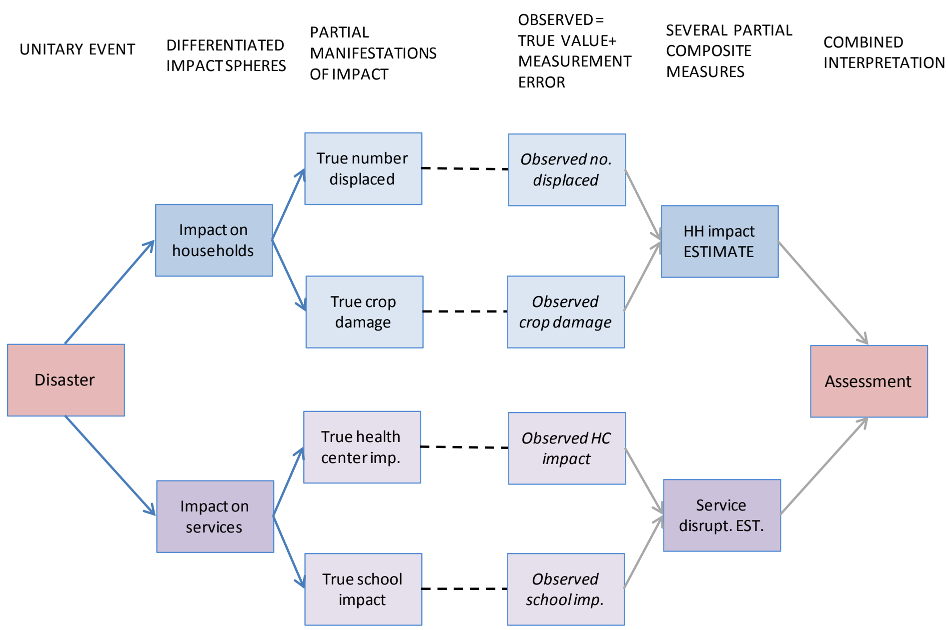
Nevertheless, in many disaster situations, *one* cause-impact nexus predominates. The impacts, however, while spawned by one major cause, travel through different institutional spheres. From the one disaster emerge multiple effects that are suffered, observed and reported in distinct conceptual fields. The distinctions arise from the speed of impacts, their relationship with human bodies and minds (notably the threat to life and dignity), from the position of those affected in the fabric of society, and last, but not least, from the institutional division of labor among responders.

What were the major pathways leading the effects of the Cambodian floods? Is it plausible to make a first distinction between impacts on households and impacts on formal organizations? Given the household-economy character of much of rural society, one is inclined to treat, as household impacts among the four impact indicators, both physical displacement and crop damage. They do not represent all the flood impacts affecting households directly, but they were the ones reported in the assessment. In severity, displacement and crop damage may be rivaled by illness and debt, but data on these, or on other plausible household impacts, were not available. In other words, what is ultimately used as indicators is a conceptual sample from a wider realm of disaster impacts.

Sampling in two dimensions

The impacts on formal organizations may have been even more variegated in type, if perhaps smaller in extent. It was the effects on health centers and schools that were reported, not those on, say, road transport or telephone service. Again, what was observed is a sampling in two dimensions: from concepts (types of impacts) and from provinces (tokens of impacts). The left half of the diagram below exemplifies these relations.

Figure 11: A conceptual approach to several partial composite measures



With the idea of institutionally differentiated impacts, we will now explore the feasibility of *separate composite measures* in depicting the diversity of disaster impacts. Two boxes on the right-hand side of the diagram call these measures, in the Cambodia flood example, "household impact estimate" and "service disruption estimate".

"Estimates" reminds us that the data on the sampled indicators represent observations, i.e. true values modified by measurement error and based on samples. This point is important in justifying separate composite measures of impacts. Without measurement error, it would be difficult to argue that "crop damage" should be lumped together with "displaced households" rather than with "schools affected". The observed correlations, as shown in the sidebar above, place crop damage closer to schools than to displacement. And sampling variance is almost absent because most affected provinces were visited.

Combining impact indicators on substantive considerations

We defy the observed correlation pattern and stick with combining "displaced households" and "crop damage" in a measure of household impact. We do so not only because uncertain measurements provide an excuse, but because loss of habitat and of crops both threaten household integrity more than disruptions in service organizations do. Moreover, from the response viewpoint, the rehabilitation of impoverished households and of damaged organizations will likely happen through distinct institutional avenues. In other words, we build composite measures on substantive grounds. We even go against the evidence (as in the observed correlation patterns) as long as probable measurement error, uncertainty from sampling or external validity aspects justify so.

Displaying results from separate measures

We calculate the estimates of "household impact" and "service disruption" as the arithmetic means of their related normalized indicators. We use the normalization developed by the UNOCHA team. We proceed so for simplicity; thus, we do not consider more demanding aggregations such as the score of the first principal component or a function that emphasizes cumulative disadvantage.

Using this procedure, we get the results shown in Figure 3 in the Summary (page 8). There appear to be four types of affected provinces, delimited by the natural break lines. Most provinces are of the "low household impact / low service disruption"-type. If the data are fairly reliable, the typology may be credible. If the errors are major and are distributed lognormally (because indicator values reflect magnitudes estimated by experts and key informants rather than exact counts), the risk of misclassification is considerable. Any of the four provinces closest to the dotted lines may in reality belong to a different impact type.

Ways to summarize

Separate impact measures let us see typologies of affected sites. With two measures, types and sites can be conveniently displayed in scatterplots. With three measures, Venn diagrams and tables can present succinct summaries, though with fewer details than in scatterplots. Summaries of more than three measures will work only in tables, but they are liable to overburden the reader. Complex tables are rarely necessary because simpler ones, as Pivot tables in spreadsheets, allow drill-down to the data table.

A tabular summary of the 16 provinces by impact type might look like this:

Table 5: Summary of affected provinces and households, by impact type

Impact type	Household impact	Service disruption	Provinces	Affected households
I.	High	High	3	87,059
II.	High	Low	2	65,948
III.	Low	High	1	72,047
IV.	Low	Low	10	124,246
Total			**16**	**349,300**

Note: Two provinces not included for lack of data.

A map showing provinces colored by impact type may usefully complement table and scatterplot, particularly if some types cluster in space. Cut-off points for high vs. low impact need to be explained in text or footnotes.

One versus several composite measures

Ordering the disaster-affected units on one composite measure has the advantage of a clear ranking. The ranking quickly identifies those in greatest need. Qualifications can be conveniently made in tables sorted by the overall rank.

There is, however, a dilemma with a global composite measure, one that collapses all the indicators:

- If they are all highly correlated, then ranking the units by just one indicator may be nearly as good and certainly will be quicker for the audience to understand. Instead of using a composite measure, one could rank and summarize by the indicator expressing the humanitarian emergency most concisely - perhaps the number of casualties, or displaced families, or - more coarsely - flooded towns and villages.
- If any of the indicators are weakly correlated with the rest, the global composite measure will obscure the diversity of situations. This diversity is likely most pronounced in the middle ranks although theoretically situations can occur in which the highest ranked on the global measure are sharply distinct on the base indicators.

The approach that we demonstrated in this section - the use of separate composite measures and the definition of impact types based on them - comes with its own set of problems. Obviously, at this level too, the indicators are normalized, weighted and aggregated, and these operations need justifying on some substantive grounds. Our formula for household impact in Cambodia, for example, implies that one percent of all cases of displacement should have the same influence as one percent of the total rice crop damaged. But separate measures have the advantage that they help create an awareness of the diversity of impacts and the distribution of affected units in that differentiated space.

Practical guidelines

We follow the ten-step procedure recommended in the OECD Handbook, although with different emphases and, for some steps, a recognition that disaster situations may affect their logic and sequence. We show alternatives and make recommendations for assessment teams working under time pressure and in rapidly evolving information environments.

1. Theoretical framework
2. Data selection
3. Imputation of missing data
4. Multivariate analysis
5. Normalization
6. Weighting and aggregation
7. Robustness and sensitivity
8. Back to the real data
9. Links to other variables
10. Presentation and visualization

Theoretical framework and data selection

Considerations

Post-disaster situation analyses are circumscribed by the scarcity of strong theories and limitations in accessing and generating data in useful time. Framework and data selection, therefore, have to be developed jointly. The framework will emerge from relevant, accessible experience with similar disasters and from knowledge of the affected area. For floods and earthquakes, the Quick Impact Analysis guidelines provide relevant checklists (Acaps 2011a, 2011b). A breakdown of impacts into pathways and/or institutional spheres can be sketched, together with the likely data sources. One can anticipate that data on more variables will become available over time, and already available data may be updated and refined. Initial proxy indicators will be replaced with more direct measures.

In the meantime, the best achievable model *within assessment deadlines* should be defined and pursued. An important consideration is for missing variables (and for missing values in otherwise measured variables) to be replaced by local expert estimates. For example, local police forces may be able to offer a magnitude estimate of newly encamped groups in influx communities long before relief agencies complete an orderly count. The police may be reachable by radio while mobile phones are still down.

Steps to follow

1. Sketch out a tentative impact tree, with the last level ("the leaves") defining indicators for which data is likely available within useful time, and the actual or likely sources.
2. Color the leaves by availability. For example: green for known to already exist and to be accessible on time; blue for needing new data collection achievable on

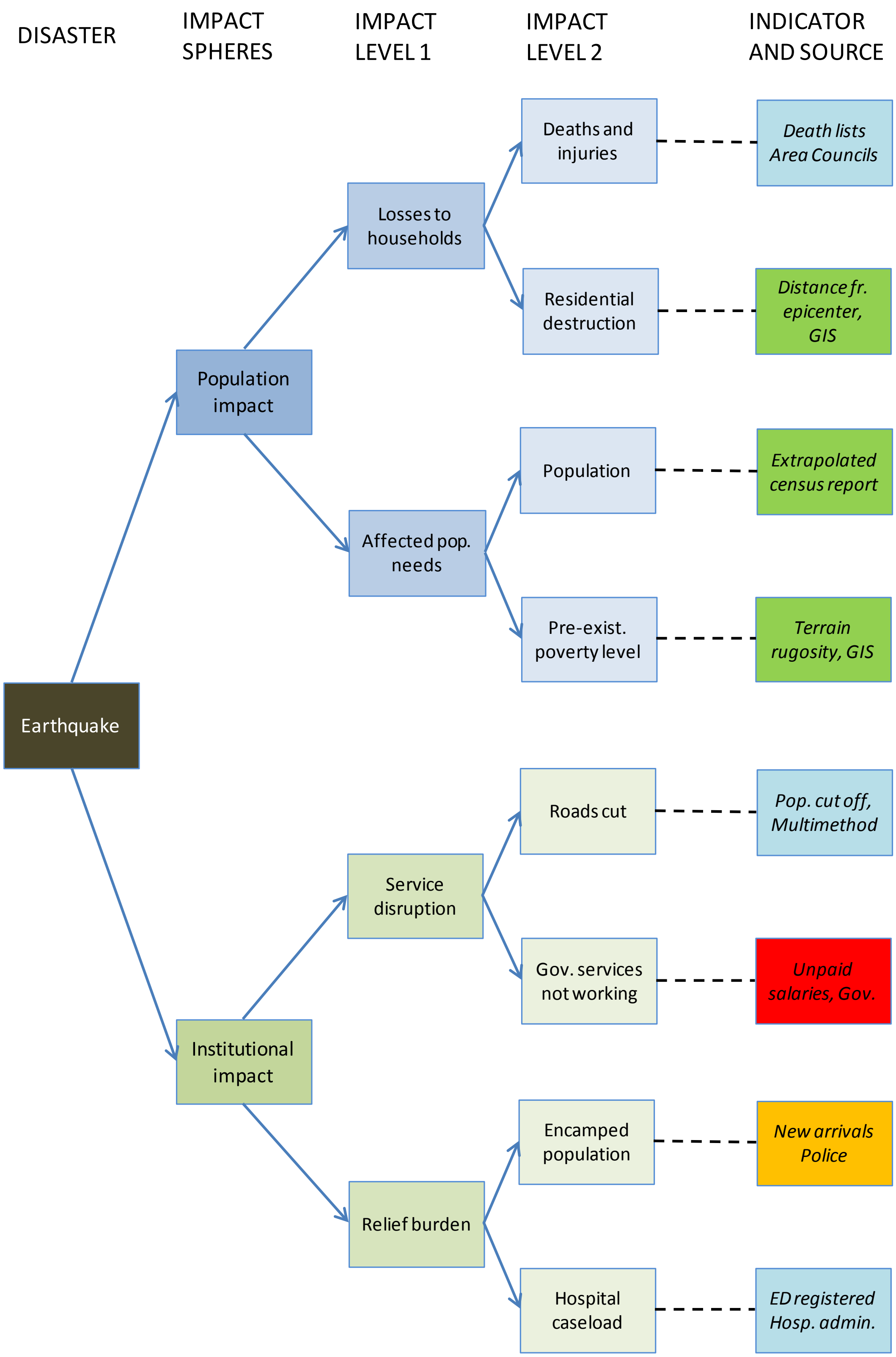
time; orange for coarsely measurable by expert opinion on time; red for important, but foreseeably too incomplete or too vague or too late.

3. Decide which of the suitable indicators can be meaningfully combined, and for what purpose: to identify the worst affected groups and areas, to anticipate priority sectors, or to anticipate the further course of the disaster.
4. Determine if they should be combined in one or in several impact measures. Should they not be combined at all, but rather be presented in a dashboard fashion?
5. Identify gaps and redundancies, chances to substitute stronger for weaker indicators, opportunity costs of additional data collections, risks of misinterpretation if those data are not collected.

[Sidebar:] Tentative impact tree for an earthquake

To illustrate point 1 and 2 of the guideline regarding theoretical framework and variable selection, we attempt this rough conceptualization for a composite measure (or measures) of the earthquake impact. See point 2 for the colors of the indicator/source boxes. The units of assessment will be local government areas, with offices and staff that have a tradition of collecting data, but may be disabled or slowed down.

Figure 12: Impact tree as a tool to balance concepts and indicators



The underlying qualitative theory is that the needs of the affected population are a composite function of direct damage to households, pre-existing conditions, disruptions of formal organizations and pressure on the remaining functioning ones. As discussed before, the indicators are a conceptual sample from the much broader range of level-2 impacts (for example, electricity and mobile phone outages could be included besides road blockages).

Note the use of proxy indicators, particularly those based on remote sensing and GIS. For example, rugosity, the "mountainness" of the terrain, proxies for poverty if mountain village and dispersed-homestead communities are poorer than valley-floor communities (Benini, Conley et al. 2009).

At this point, nothing is said yet about necessary indicator transformations, the expected correlation structure and whether ultimately one, several or no composite measure will be constructed. The discussion focuses on plausible models: impact types, indicators expressing their breadth and coherence, data availability and assessment deadlines.

Imputation of missing data

Considerations

Missing values in impact indicators are likely not "completely missing at random":

- Assessment teams may find information access degraded apace with severity. Physical access or security may be badly compromised in the worst affected areas. Their key informants may have died or moved.
- Conversely, when accessibility is not a major problem, authorities may allocate assessment resources chiefly to those areas which first reports indicate are the worst affected, to the temporary neglect of reportedly less affected ones.

Either situation, technically speaking, causes missing values in the indicator variables that correlate with other observed variables or with the underlying severity.

Three strategies to fill missing values or to exclude and footnote cases seem feasible:

- If sufficient comparability exists on the observed indicators, missing values may be filled in with a representative value (mean, median, minimum) for the comparable subgroup. For example, a subdistrict without death figures could be given an estimate from the deaths suffered in neighboring subdistricts (it may have to be population-weighted). This should be attempted only if such cases are few, comparables are easily discerned, and the operation takes little time.
- If that is not feasible, yet most of the other indicator values for the case in point do exist, then an average for the entire affected population may fill in. Since impact distributions are often skewed, the median is preferred. Alternatively, an expert or key informant - such as a local staff member from the concerned area - can be asked for an estimate if we believe that this person knows.

- If the case has blanks on most indicators, it may be better to exclude it from the composite measure calculations. This must be footnoted, and if updates fill the missing information, such cases should be re-included.

More advanced statistical imputation methods do not seem warranted, notably because samples tend to be purposive, and many missing values may be filled in when updates arrive from the field. Documenting the fill-in steps, including time-stamping the data table version, is important.

Steps to follow

1. In the indicator data table, mark missing value cells (and cells with unusable values - hopeless outliers, ranges instead of point values, wrong units [households were population was expected], etc.) in a particular color. In a large data table, name a column on the right side as "NoMV" (number of missing values) and use formulas of the kind

   = COLUMNS(RCxx:RCyy) - COUNTIF(RCxx:RCyy, ">=0")

   for row totals of missing numeric values and values < 0, where the mixed reference *RCxx:RCyy* denotes the range of indicator values for the case. Sorting by, or heat-mapping, this auxiliary variable helps discern patterns[11].
2. Study the pattern of missing and inappropriate values for each indicator as well as jointly.
3. Decide an imputation strategy on the considerations outlined above. Balance loss of cases (if you did not fill in), effort and potential for misunderstanding.
4. For imputations, always create a new field, such as "PopRevised[Date]" for "Population". Never overwrite blanks or invalids in the original data column, except for unambiguous data cleaning. Explain complex changes in a comment field, not in cell comments (which will never be read by other users).
5. If you wish to exclude cases from subsequent analyses, create a tagging variable, such as "Include", with 1 for cases retained, 0 for those to exclude. Do not delete original records.
6. In a large data table, in order to replace missing values in the revised indicator field, use a formula for uniform replacements.

For example,

= IF(ISNUMBER(RCoriginal) = TRUE, RCoriginal,
IF(RCnomv > 3,"", MEDIAN(x-indicator_range)))

where *RCoriginal* points to the original indicator cell for the case, *RCnomv* points to the cell holding the number of missing values, *x-indicator_range* is the named column vector of the original indicator. The condition *IF(RCnomv > 3,"", MEDIAN*.. tells Excel not to

[11] The same can be done columnwise, as = COLUMNS(RxxC:RyyC) - COUNTIF(RxxC:RyyC, ">=0") for the number of missing and negative numeric values per indicator, but these formulas should be removed soon so that they are not inadvertently included as spurious record.

fill in for cases with more than three indicators with missing values (if this is your policy), else to fill in the median of the observed values.

[Sidebar:] Impact rank stability when missing values are imputed

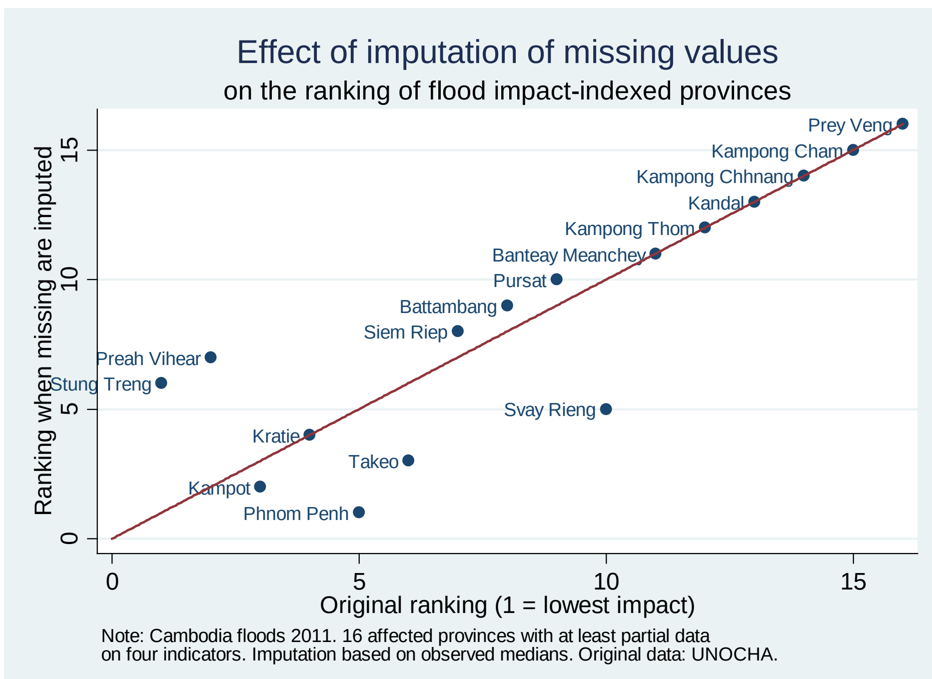
The Cambodia flood data invite a small experiment. Only nine of the 18 affected provinces have complete indicator values (and even this is true only if we disregard the missing in the school indicators from which the school needs subindex was formed). Two of the provinces in the original data table are virtually without information. They are excluded from the analysis, and also from the imputation of missing.

Recall that the UNOCHA team used an implicit imputation method: their aggregation formula treated missing values as zero.

The question of interest then is this: If we use a different imputation formula, how will the ranking based on the recalculated impact index compare to the original ranking?

We demonstrate the mechanics of using a median-based formula in the sheet "C_MissingImputed" in the demo workbook.

Figure 13: Rank changes as a result of imputing missing indicator values



As one would expect, imputing by medians tends to elevate provinces that ranked very low in the original UNOCHA formula AND had missing values (Stung Treng and Preah Vihear). It penalizes, because of the re-normalizing of the indicators with higher totals, provinces that had complete observations (most drastically Svay Rieng). However, the six most-impacted provinces are so robustly above the rest that the re-normalizing does not affect their relative ranking.

One cannot generalize much from this example. In Cambodia, for reasons that we do not know, impact indicators have missing values primarily in provinces with lower numbers of affected households. This is noteworthy because it is contrary to the expectation that reporting would be least complete from the most affected areas. One may speculate that crop damage and school needs assessment resources were directed with priority to provinces considered highly affected in early reports.

Multivariate analysis

Considerations

The major attraction of composite measures in rapid needs assessments arises from the fact that disaster impacts are correlated. The designer of such measures thus needs to understand the correlations among the candidate indicators. However, before the chosen indicators are aggregated into a composite, they need to be normalized and weighted. The normalization (and also aggregation) choices require that the basic descriptive statistics be computed, inspected, and considered for their implications. Therefore, the simple adoption, from the OECD Handbook, of "multivariate analysis" at the point of the design process is incomplete, to say the least. There is no way to bypass the mundane work of data inspection in initially purely descriptive ways.

A second consideration has to do with speed and efficiency. With increasing numbers of candidate indicators, adequate initial data exploration using a spreadsheet application grows tedious and prohibitively time-consuming. There are features and add-ins to Excel that considerably accelerate the production of descriptive statistics. Naming the column ranges that hold the indicator data, various elements of the Data Analysis Tools, the function INDIRECT used to create basic descriptive statistics in a metadata table alongside the names of the variables, etc. stretch the range of procedures that one can do somewhat efficiently under time pressure. We highly recommend *SSC-Stat*, a free Excel add-in offered by Reading University[12]. This tool offers rapid descriptive statistics as well as a host of other applications in data manipulation, visualization and analysis.

At this point the advantages of statistical programs come to outdo the costs of specialist expertise, reduced sharing and communication difficulties that militate against their use for many common analysis needs in rapid assessments. Also, programs such as STATA, SPSS and the freeware R move seamlessly from basic descriptive statistics to multivariate analysis. Some of the analysis types helpful in exploring the internal structure of composite measures - rank order correlations, factor and cluster analysis - can be performed in Excel only with the help of high-end, fairly expensive add-ons. Many ordinary descriptive tools - histograms, correlation matrices - are simply a lot faster in the statistical applications.

Our gut feeling is that whenever more than four candidate indicators are to be investigated jointly, it is time to supplement the spreadsheet application with a proper statistics program.

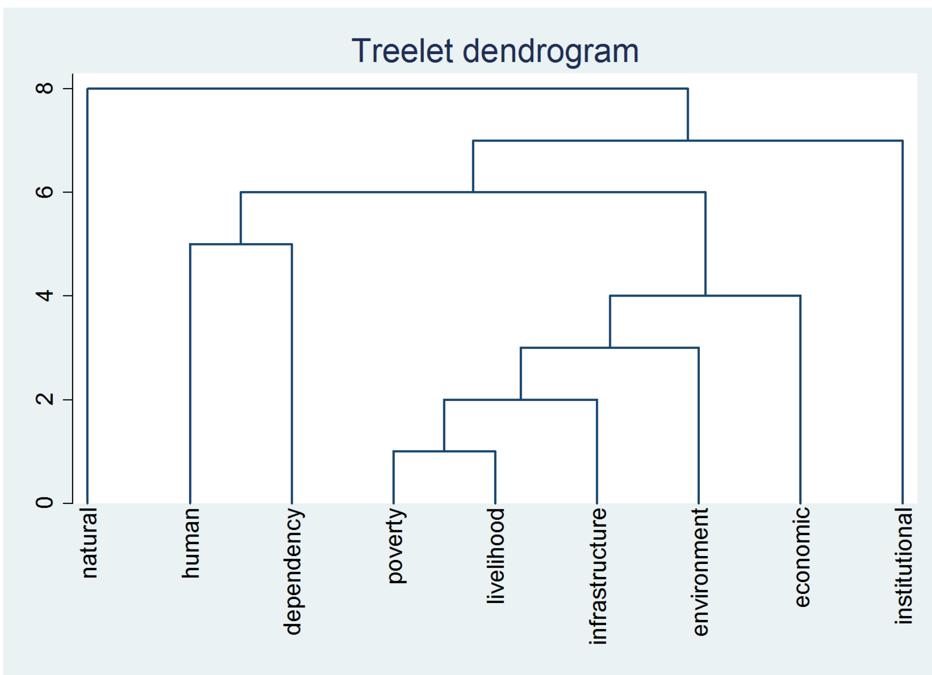
[12] Downloadable from http://www.reading.ac.uk/ssc/n/software/sscstat/helpfile/ht_start.htm.

[Sidebar:] Speedy insight into the indicator correlations

The OECD Handbook correctly emphasizes the need for *"data reduction and synthesis, simultaneously in the direction of objects and variables"* (op.cit., page 27). The challenge is tricky because separate factor and cluster analyses can mis-inform each other and the analyst: Objects (the assessed sites) fall into common clusters because they have similar positions on several variables (indicators); variables that should be excluded as redundant are retained because clustered objects cause strong correlations.

The Handbook recommends a procedure - factorial k-means analysis (Vichi and Kiers 2001) - promising to overcome the dilemma. Until recently a commonly accessible implementation did not exist. With the recent *treelet* transform introduced in STATA (Gorst-Rasmussen 2012), a similar tool has been put at the fingertips of composite measure designers. We demonstrate its power with a dendrogram of the nine indicators flowing into the Global Focus Model. Indicators whose branches meet at a lower level are more closely related to each other (there is, of course, more to it!).

Figure 14: Fast and crisp representation of indicator correlations



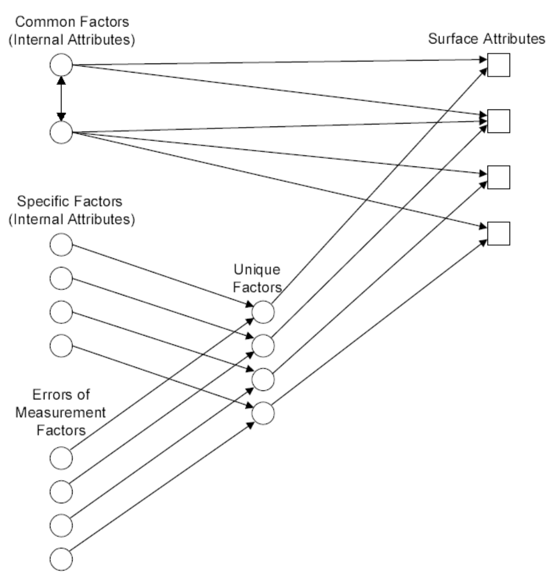
The workup to this result may have taken one minute. The diagram immediately reinforces the intuition that "poverty" and "livelihood" may be substantively redundant, and this redundancy would have to be controlled (as it is, through the GFM weighting scheme). The visual relationship may also raise doubts about the wisdom of lumping natural hazards and institutional capacity into one composite risk measure. These doubts would have to be spelled out in a detailed statistical and policy analysis.

Biplots, as demonstrated further above, present cases and variables in a unified visual format. They too can be rapidly produced - and thrown away if we are not satisfied!

Steps to follow

1. At the minimum, inspect candidate indicators in their univariate (histograms) and in some (scatterplots in Excel, as time allows), ideally all (in matrix plots in a statistical application) bivariate distributions. Table basic descriptive statistics, evaluate outliers for plausibility and substantive implications (e.g, characterizing a region as being hit by an extraordinary catastrophe).
2. If outliers are dubious, create a new variable in which you replace them with expert estimates, with a protocol note.
3. Create correlation tables. Up to four indicators (for some people, more), visual inspection can recognize patterns. Re-inspect unexpectedly high or low correlations in scatterplots. If limited to spreadsheets, go to point 6.
4. With more than four indicators, use a statistical program to produce a matrix plot. Produce both rank-order and Pearson moment correlations. Compare the two. If differences are minor, proceed to an exploratory factor[13] or treelet analysis. If they are major, transform the indicators with major differences into normal scores[14]; use these variables for factor or treelet analysis, and check whether the suggested associations make sense when looking at the untransformed scatterplots.

Figure 15: Basic factor-analytic scheme



Source: Tucker and MacCallum (1997: 7). "Surface Attributes" correspond to our indicators. Note that the measurement errors in this scheme are not correlated. This assumption is unrealistic for rapid needs assessment. We must expect errors to contribute to the inferred factors.

5. Retain a solution with only factors with eigenvalue > 0. Note what proportion of the common variance the first factor accounts for. Try for a meaningful interpretation of the loading pattern, at least in the first factor (perhaps after rotation). Run again for one factor only, to get the uniqueness of each indicator. Those with uniqueness < 0.60 may be unproblematic to include in the composite measure (unless they are loading

[13] The Handbook is indifferent to the choice of principal component vs. factor analysis. Since our objective is to predict the correlation structure with as few factors as possible (not: to exhaust the total variance of the indicators, as PCA attempts), one is inclined to prefer factor analysis. However, analysts should have their own or get a second opinion on this point. In practice, results seldom differ greatly. Treelet analysis works with PCA and elements of cluster analysis.

[14] In the second of the two meanings given in the Wikipedia article (Wikipedia 2012b), *"assigning alternative values to data points within a dataset, with the broad intention of creating data values than can be interpreted as being approximations for values that might have been observed had the data arisen from a standard normal distribution"*.

highly because of other substantively redundant indicators included). More unique variables can be included if there are substantive (policy!) reasons justifying it. - For treelet analysis, vary the cut level on the cluster tree, interpret the sparse first component, and use the cross-validation feature to suggest an optimal cut level. However, evaluate indicator inclusion and exclusion on substantive grounds.

6. Determine which indicators to finally retain in a common composite measure. Substantively redundant indicators may be combined in a subindex, or they are included after normalization with their weights adjusted.

Normalization

Considerations

Normalization is the first of three necessary transformations, followed by weighting and aggregation[15]. Normalization produces standardized measurement units. Theoretically, the weighting system could take over this part too (as in price systems), but it would then lose its function to signal easily recognized importance. The Handbook discusses a number of normalization methods (op.cit: 27-31). Other methods too have been used for composite measures: linear normalization (dividing by the maximum value), vector normalization (dividing by $[\text{sum}_i(x_{ij})]^{-0.5}$) (Yoon and Hwang 1995: op.cit., 16), normal scores (mentioned above, for factor analysis purposes).

The UNOCHA team in Cambodia normalized the indicators by dividing them by their column sums, in other words, for each indicator i calculate $[\text{sum}_i(x_{ij})]^{-1}$. This approach has a number of attractive properties:

- The normalized values are identical to the original except for a scaling factor.
- The transformation is intuitively understandable also to all audiences.
- If the fractions in the indicator sum are expressed as percentages, they may be understood as proportionate to cases concerned and even as suggestive of caseloads awaiting the responders.

Users of this type of normalization should, however, keep in mind that the results are distribution-dependent. See the example above, in Table 3 on page 29. This has to be reflected in the choice of weights. Using percentages to make impact distributions more intuitive is *formally* problematic if the units are only a small sample of all units affected. It would be *substantively* problematic if understood directly as some relief allocation model.

In some quarters min-max normalization is popular. Min-max creates an identical range [0, 1] for all indicators, by subtracting the lowest value and then dividing by the range. Yet, dividing by column sum seems preferable to min-max normalization, at least in the humanitarian realm. If, in an indicator "persons killed in violence", the lowest observed value is 10, min-max would produce a normalized zero. This could be mis-interpreted as

[15] This does not hold absolutely for so-called non-compensatory methods of preference formation, with which this note deals only in a limited way. Some of these methods do not use weighting; and "aggregation" in their context is a misnomer for other operations.

a unit not affected by deadly violence. Even if that were avoided, the min-max formula disproportionately discounts low values.

Steps to follow

1. Normalize indicators by dividing each of them by the sum of its values.
2. Express the normalized values as percentages only if percentages are unlikely to be misinterpreted.
3. Consider the distribution dependence of the normalized values during the weighting process.

This is most easily done by comparing the histograms of the normalized indicators side by side. Comparing medians and skews can also help.

4. Consider allowing a higher weight for an indicator on which many units scored high and for which (as a result of dividing by a larger sum) the high-end values were relatively more scaled down than high-end values in other indicators.

Such adjustments have to be done on *substantive* grounds, in the sense of "Given the distribution of the original values, this indicator deserves, for such and such policy reason, a higher weight than the originally intended weight would express as its importance."

Weighting indicators

Considerations

The weighting of indicators is the vulnerable open flank through which critics attack composite measures. The measures are considered of questionable validity because the weights are essentially arbitrary. The arbitrariness charge is leveled at all composite measures except those that rely 1. on natural weights (such as the market prices in the economy) or 2. on weights that were calibrated through previous research into the connection between indicators and some key outcome variable[16]. The charge sticks also to measures that do not weight indicators explicitly. These use identical weights of one, so-called "unit weights".

Although we will wind up recommending simple methods, we need to dwell on the weighting questions in some depth. This is where composite measures are most vehemently attacked; the analyst should know some of the issues and alternatives:

Do weights express importance? - The OECD Handbook is highly critical of compensatory methods - methods in which a gain in one indicator is offset by a loss in any others, and vice versa, and where the extent of compensation depends on the relative weights. Because of the compensatory mechanism, the Handbook denies the weights the status of true "importance coefficients." (op.cit.: 31 et passim).

---

[16] An example of the latter would be laboratory studies in which the contributions of various chemical pollutants to ozone destruction are determined, and these parameters are then used in environmental harm assessments looking at various industries.

This criticism seems overwrought. Surely, if we double the weight on indicator X, we double its importance relatively to what it was in the previous weights vector. Weights do express importance.

Should local people set the weigths? - The Handbook is easier to agree with when it recommends considering other weighting schemes. For example,

> *"participatory methods that incorporate various stakeholders – experts, citizens and politicians – can be used to assign weights. .. In the budget allocation approach, experts are given a "budget" of N points, to be distributed over a number of individual indicators, "paying" more for those indicators whose importance they want to stress .. . The budget allocation is optimal for a maximum of 10-12 indicators"* (p.32).

Such weights could be established also from the key informant-based surveys of affected communities and could perhaps be inferred from Borda counts of sector priorities (Lansdowne 1996; Wikipedia 2011), to the extent that these speak clearly to the indicators.

Are the weights already in the data? - There are other ways for the analyst to eschew the responsibility of setting weights, justifying them and dealing with possible dissension. One is to simply let the data speak for themselves. The so-called "Benefit of the Doubt" method (Cherchye, Moesen et al. 2007) assigns different weights for different units. It does so in a way that lets each unit maximize its composite measure, relative to the values of the composite that the other units gain using these particular weights.

We gave a simple example of this principle in Figure 4 on page 16. Whatever the weights chosen on the two indicators, C' will always be the highest-ranking on the composite. Allowing different weights for A' and B', their composite scores relative to C' will be maximized by setting the weight on the indicator on which they score higher to one, and the other weight to zero. Since A's number of displaced households is around 80 percent of C's, and B's number of school closures is of similar proportion, A' and B' can be said to be similarly impacted - given their different indicator weights![17]

Is there anything better than unit weights? - Some authors have come out in defense of simple additive weighting *and* of using unit weights for the purpose (Bobko, Roth et al. 2007). Their main argument is that the ranking of units did not change significantly when weights determined by other methods were used instead of unit weights. We made the same finding when varying the weight for one of the four indicators in the Cambodia floods assessment (see above, Table 4 on page 31). Only the ranks of some of the original midfield units changed substantially, yet without propelling any of them to top or bottom ranks.

[17] We have applied this method - which is based on Data Envelopment Analysis algorithms - to the Global Focus Model data. We tested how much the country ranks change over the original ranks. The changes were found to be minor. Given the technical difficulties of the method, we recommend it only for analysts equipped to use the underlying statistical model and confident to explain the results.

We have reservations against unit weights unless they are justified on substantive grounds. Not everything is equally important in assessing disaster impact. Weights should be used to deliberately express differences in importance. Only if the differences are judged to be minimal, or if different weights will likely provoke confusion and dissension, should unit weights be considered, as the lesser evil.

One more observation is due here. The Handbook stresses as one of the weighting challenges the problem of redundancy, which it links to the danger of double-counting. This problem is real, but we do not follow the Handbook's formulaic solution of assuming that any two indicators correlated above a defined threshold (e.g. a Pearson correlation coefficient of 0.80) are suspect of double-counting (p. 32). We believe that adjustments for double-counting (either by elimination of one of the indicators, or by halving their weights) should be done for *substantively* redundant indicators. "Income poverty" and "livelihoods" are substantively redundant, "income poverty" and "child malnutrition", though highly correlated, are not. The redundancy should be investigated, and adjustments made, during the initial indicator selection and multivariate analysis. It is not primarily a problem of weighting.

What can we recommend practically, in view of this host of (sometimes conflicting) considerations?

- For **simplicity**, the UNOCHA Cambodia assessment example - simple unit weights on normalized indicators - is compelling. Unit weights are recommended also if unequal weights of any kind would not find consensus among stakeholders.
- From a **policy** viewpoint, weights should be deliberately set to express differences in importance. In the Cambodia example, displacement and crop loss are likely, in the eyes of many observers, to merit a higher weight than school needs, with health center closures falling somewhere in-between.
- In situations where any worsening in **a supremely important indicator** cannot be offset by any amount of gains in other indicators, non-compensatory methods should be attempted. In the Japanese "triple disaster", for example, evacuees cannot return to high-radiation zones even if their homes suffered minimal or no earthquake and tsunami damage. The latter cannot compensate for high nuclear radiation exposure.

The sidebar below demonstrates a possible non-compensatory method feasible for spreadsheets. It is an approximate method developed for simplicity and speed and possibly less valid than the sophisticated, but non-implemented methods proposed by the Handbook.

Steps to follow

1. Determine whether any of the indicators is clearly more important than the others *in all circumstances* for the units of this needs assessment.
2. If so, find a suitable non-compensatory method (e.g., the one demonstrated below). Else use a compensatory method (weighted indicators):

3. If some indicators should receive higher importance *and* stakeholders can agree on importance differences, set different weights. Higher importance can be given also to adjust for inappropriate normalization effects, as argued further above.
4. Else use unit weights on the normalized indicators.

[Sidebar:] An example of a non-compensatory method

As noted at several points, compensatory methods are those that permit a loss in some indicator to be compensated by gains in one or more other indicators, and vice versa. Where such compensation is not appropriate - see the argument above from the Japanese triple disaster -, non-compensatory methods are called for. These methods do not per se produce composite measures. Rather they establish preference orders. "Non-compensatory" means that inferiority in a more important indicator cannot be offset by superiority in less important ones. A proxy composite may be derived as an ordinal variable from the overall ordering of the units.

There are several non-compensatory methods in the decision sciences (Yoon and Hwang 1995: 17-31, for a quick, if outdated overview). We exemplify them here with the *"Lexicographic Semi-order Method"* (Luce 1956; Tversky 1969; Manzini and Mariotti 2012). Lexicographic methods in decision-making sort alternatives strictly by the most important attribute and consider lesser attributes only when alternatives are tied on the more important ones. A semiorder is an ordering of items in which two items are considered incomparable when the difference between their scores on the ordering variable is smaller than some threshold of distinction (Wikipedia 2012c).

We choose this method because it works with relatively simple spreadsheet operations. We apply it to the Cambodia data. The basic idea is that decision criteria (in our case: indicators) are ordered by importance, but the order is not absolute. It is weakened by the recognition that small differences in a more important indicator should not overrule large differences in the next less important indicator.

This weak ordering principle may be adopted on substantive grounds (meaning: the difference in the importance between two indicators may be limited, and higher importance cannot be endlessly stretched). Alternatively, it may be justified because of suspected substantial measurement error. In other words, an exaggerated value on the more important indicator should not be allowed to discount another unit with the same true value and with a higher value on the next less important indicator. In rapid needs assessments after the disaster, when information is noisy, both motivations mix.

Let us now assume this importance order for the four indicators used by the UNOCHA team in Cambodia:

household displacement > crop damage > health centers affected > school needs,

where *a* > *b* means that a is more important than b. Further, we assume that, for the sake of example, we are primarily interested in measurement error. We assume that the error is proportionate to the observed value. We are ready to give a unit value the same rank as the units with the higher values on the indicator in point if the difference is not more than 20 percent.

For convenience we name the normalized indicator ranges, as e.g. "Displace", "Crops", etc.

To compute the semi-order, we first sort, ascendingly, on the most important indicator, the shares of displaced household. In the next column to its right, we calculate the roundup for units whose values are within a 20 percent reach of the next (one or several) higher values.

A suitable formula accomplishing this is

*=VLOOKUP(1.2 * RC[-1],Displace,1,TRUE)*

It adds 20 percent to the lookup value, searches in the range called "Displace" (which trivially has only one column), makes this search approximate (TRUE), and finds the nearest value equal or smaller than 1.2 * RC[-1]. This table shows the result, for this first indicator, with values of provinces that have been "upgraded" framed with a red border.

Table 6: Adjustment of values in the lexicographic semi-order method

Provinces	ShareDisplaced orig.	ShareDisplaced (adjusted if diff. =< 20%)	ShareRiceDamag orig.	ShareRiceDamag (adjusted if diff. =< 20%)	ShareHCAffect orig	ShareHCAffect (adjusted if diff. =< 20%)	School needs indic orig.	School needs indic (adjusted if diff. =< 20%)
Siem Riep	0.000	0.000	0.092		0.067		0.008	
Stung Treng	0.004	0.004	0.000		0.011		0.000	
Preah Vihear	0.013	0.015	0.000		0.016		0.000	
Takeo	0.014	0.015	0.034		0.032		0.080	
Kampot	0.015	0.015	0.000		0.011		0.013	
Battambang	0.023	0.027	0.000		0.056		0.089	
Kratie	0.027	0.031	0.000		0.043		0.014	
Pursat	0.031	0.031	0.000		0.037		0.106	
Kandal	0.042	0.047	0.035		0.152		0.150	
Kampong Thom	0.047	0.047	0.175		0.035		0.103	
Phnom Penh	0.058	0.058	0.004		0.019		0.005	
Svay Rieng	0.081	0.081	0.047		0.037		0.013	
Banteay Meanchey	0.104	0.118	0.115		0.123		0.010	
Kampong Cham	0.118	0.118	0.139		0.144		0.201	
Prey Veng	0.198	0.224	0.289		0.163		0.100	
Kampong Chhnang	0.224	0.224	0.068		0.051		0.106	

We replace the VLOOKUP formulas with their results (copy - paste, values only) so that we can sort on the next indicator, crop damage. We use this function again and repeat the operation through all indicators. Finally we sort hierarchically on the adjusted indicators. The table, sorted on the semi-order (rightmost column), presents as follows.

Table 7: Results of the lexicographic semi-ordering

	Indicators and adjusted indicators								Ranking schemes		
Provinces	ShareDisplaced orig.	ShareDisplaced (adjusted if diff. =< 20%)	ShareRiceDamag orig.	ShareRiceDamag (adjusted if diff. =< 20%)	ShareHCAffect orig	ShareHCAffect (adjusted if diff. =< 20%)	School needs indic orig.	School needs indic (adjusted if diff. =< 20%)	Rank of the UNOCHA composite	Strict lexicographic order	Lexicographic semi-order
Siem Riep	0.000	0.000	0.092	0.092	0.067	0.067	0.008	0.010	7	1	1
Stung Treng	0.004	0.004	0.000	0.000	0.011	0.011	0.000	0.000	1	2	2
Kampot	0.015	0.015	0.000	0.000	0.011	0.011	0.013	0.014	3	5	3
Preah Vihear	0.013	0.015	0.000	0.000	0.016	0.019	0.000	0.000	2	3	4
Takeo	0.014	0.015	0.034	0.035	0.032	0.037	0.080	0.089	6	4	5
Battambang	0.023	0.027	0.000	0.000	0.056	0.067	0.089	0.106	8	6	6
Pursat	0.031	0.031	0.000	0.000	0.037	0.043	0.106	0.106	9	8	7
Kratie	0.027	0.031	0.000	0.000	0.043	0.051	0.014	0.014	4	7	8
Kandal	0.042	0.047	0.035	0.035	0.152	0.163	0.150	0.150	13	9	9
Kampong Thom	0.047	0.047	0.175	0.175	0.035	0.037	0.103	0.106	12	10	10
Phnom Penh	0.058	0.058	0.004	0.004	0.019	0.019	0.005	0.005	5	11	11
Svay Rieng	0.081	0.081	0.047	0.047	0.037	0.043	0.013	0.014	10	12	12
Banteay Meanchey	0.104	0.118	0.115	0.139	0.123	0.144	0.010	0.010	11	13	13
Kampong Cham	0.118	0.118	0.139	0.139	0.144	0.163	0.201	0.201	15	14	14
Kampong Chhnang	0.224	0.224	0.068	0.068	0.051	0.056	0.106	0.106	14	16	15
Prey Veng	0.198	0.224	0.289	0.289	0.163	0.163	0.100	0.106	16	15	16

Red-bordered cells mark all the cells that were adjusted upward while the table was sorted on their particular indicator. The four white cells in the share-of-displaced-households column mark cases where the semi-order deviates from the strict order (because of higher values on the second and/or third indicators).

The result confirms the robustness of rankings across multiple methods. First, compared to the strict ordering by most important and less important indicators, seven rank changes are noted. All are minor, and all are informed by the adjusted values of the three most important indicators - displacement, crop loss, and health centers. School needs are irrelevant under this adjustment.

Second, compared to the original ranking (based on the simple additive unit weighting, i.e. a compensatory method), only two provinces have changed rank substantially - Siem Riep, which is now considered the lowest-impacted, and Phnom Penh, which shot up from rank 5 to rank 11.

Thus, the changes are minor. A higher adjustment factor (say 50 instead of 20 percent) might have created more equally valued units on the first two or three important indicators and would thus have led to more reversals informed by lower-importance indicators. But it is doubtful that this would modify the overall ranking dramatically (for time constraints, we have not done this).

The insignificance of the rank changes suggests that the greater analytic effort required to produce a lexicographic semi-order, compared to the easy unit weight-based compensatory method, may in many or most situations not be worthwhile. Other non-compensatory methods may offer less flexibility than the semi-order method or may be too difficult to compute in a spreadsheet. However, the exercise demonstrates the basic difference between compensatory and non-compensatory methods.

Aggregation

Considerations

Aggregation is the operation that combines the normalized and weighted indicators into the composite measure. Again, different methods are in practice. Linear aggregation means that the indicator values are simply added. Geometric aggregation multiplies the indicators. By definition, if a unit, in any of its indicators, scores zero, its geometric composite is zero. Aggregation in the case of non-compensatory methods may be used for lack of a better term; it combines the several operations that produce an ordinal ranking.

Since it is quite common that units affected by a disaster score zero on some of the indicators, geometric aggregation is generally not suitable for rapid needs assessments.

By the time we aggregate the common assumption is that substantive questions have been decided. They were settled primarily by the grouping of base indicators into sub-indices (as, e.g., for hazard, vulnerability and capacity in the Global Focus Model) and by the weights assigned to base indicators and sub-indices. In most cases, therefore, aggregation should proceed almost unnoticed, as simple addition.

Rarely, situations may occur that advocate for more complex aggregation functions. Such merit consideration when significant non-linear cumulative effects are expected from various partial effects (expressed in the indicators) on the total effect (e.g., the country risk in the Global Focus Model, or the impact on communities in a post-disaster assessment). A family of aggregation functions that handles cumulative effects is described in the appendix and implemented in the demo Excel workbook.

These functions have a pedigree both in mathematics and in the Human Development Index design. However, the conceptual and computational effort is considerable, and the composite measure thus constructed will not be intuitively understood. Using such an alternative aggregation function with the Global Focus Model data did not significantly change the country ranking. While the analyst might want to know that there are alternatives, we advise to stick to the simple additive formula.

Steps to follow

1. Aggregate the normalized and weighted indicators by simple addition.
2. In non-compensatory methods, aggregation is a misnomer. It amounts to the operations that produce the final ranking or ordinal preference variable.

Testing the composite measure

Robustness

Considerations

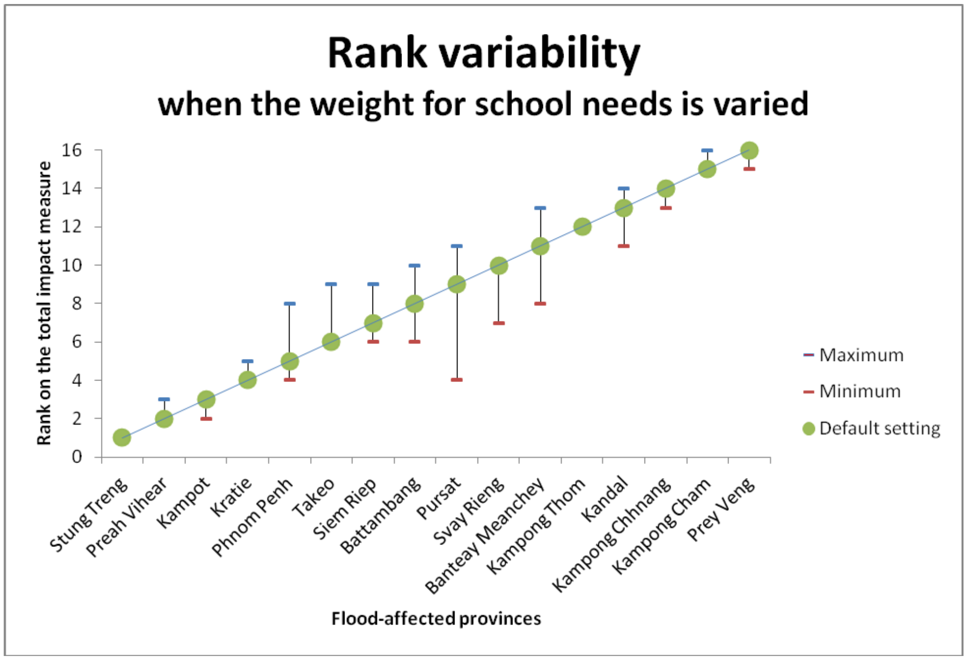
The OECD Handbook deals with robustness and sensitivity at length and with sophisticated statistical methods. All the same, its definitions are not unambiguous. Here we define *robustness* as a deterministic property of a composite measure. By contrast, *sensitivity* summarizes the influence of stochastic processes. A composite is robust if changes in a known parameter - particularly the indicator weights - have but a small effect on the ranking of units. The composite is sensitive to the extent that uncertainty in an unknown parameter (such as a weight to be determined by a regression analysis) or in a particular indicator, from sampling variance and measurement error, translates into variable rankings. To illustrate: the choice of the normalization method determines composite outcomes that are more *robust* or less so. The sample dependence of the normalized indicator values adds to the *sensitivity* of the measure.

In needs assessments, time pressure and toolbox narrowly circumscribe how much can be done in terms of robustness as well as sensitivity analysis. As for robustness, it appears fair to expect that the analyst inspect the effect of changing indicator weights. This can be done with more or with less effort, dictated by timing, skill level and amount of data.

For small tables, with fewer than twenty units, manual changes in one weight at a time and visual inspection of the ensuing rankings may suffice. The rankings are most easily surveyed by placing a static copy of the ranks under the default weights next to the column in which the varied results appear. To speed up the work, and have a sure grasp at every moment of what the weights are, they should be named (i.e., as Excel named ranges) in a small external weights table and referenced by name in the formulas of the composite measure table.

For analysts familiar with the Excel function TABLE, a more complete and visually compelling approach is feasible. This is demonstrated in the appendix; results were already presented in the Cambodia floods assessment section. The robustness can be visualized in variability charts (in Excel: high-low-close charts), but in most rapid assessment situations this will be a time-consuming luxury.

Figure 16: Robustness-to-weight-changes chart



The chart once again confirms what we already know theoretically and from the results of the tabular simulation: that the ranks are the least robust in the mid-field. In other words, the units most impacted by the disaster are likely to be in the top ranks under all or a wide variety of weight settings.

Steps to follow

1. Normally, limit robustness tests to variations in indicator weights. Extend them to other elements of the composite measure (normalization method, aggregation function) only if there are compelling reasons, such as a dispute over methods.
2. Depending on time, number of units and skills, limit the tests to a small number of manual changes with visual inspection, or use the TABLE function for a more extensive study.
3. Inspect closely the stability of ranks among the units shown most impacted under the default weights setting (e.g. the unit weights used in Cambodia). If changing the weight of some indicator within a sensible range produces dramatic rank changes, several possibilities should be considered:
   - Exclude this indicator if it is not of critical importance
   - Abandon the one all-encompassing measure ambition. Instead cross-tabulate this indicator against others or against some reduced composite.
   - Check for errors in data or formulas.

Sensitivity

For definitions vis-à-vis robustness, see above.

The academic approach is to simulate indicator values drawn from plausible distributions and covariances. This emphasizes the *substantive* dimension of the composite. If we vary also the units, by repeatedly sampling from a population, we add a *social* dimension to sensitivity tests.

However, in rapid needs assessments, the leading dimension is *temporal*. To a lesser degree, it is social. Indicator values are updated repeatedly, from broad first indications to more reliable and precise estimates and later to counts and measures from surveys. Samples are extended - i.e., more sites are assessed - or may even end in a full census of the units of interest.

The major challenge regarding sensitivity, therefore, is not so much one of data *analysis*, but of data *management*. There will be a sequence of data tables, each one partially updated and extended from the previous tables. As data arrives in finer grain, units will have to be split, moving from provinces to districts to communes. The refinements may not be provided for all indicators - there may, for example, be district-wise counts of damaged health centers, and estimates of displaced households only by province.

Designers of composite impact measure will therefore have to make opportunistic choices and be ready to revise their work multiple times. A suitable administrative level at which to calculate the measure has to be chosen. A higher level (e.g., province) loses information by aggregating indicators for which data exists at the lower level (e.g., district). Conversely, a lower level poses problems of attribution and apportionment (e.g., on a pre-disaster population basis) when merging higher-level data downwards.

The particular data management challenges will vary from disaster to disaster. Rules beyond generic data management practices are hard to devise in the abstract. The assumption, though, is that over time information will grow more precise and more finely grained. Testing for the effects of uncertainty in the data boils down to "wait and see" until the assessment receives or generates more and better data, and then incorporating the new data in revised versions of the composite measure. In a more proactive stance, if the analyst has the time and the guts to second-guess the data, the leverage of dubious inliers or outliers can be investigated. But how would he communicate such experiments to other team members and beyond?

This may appear fatalistic, but given all the known limitations, an updating strategy, with proper documentation for successors and consumers, is the best that may be expected of the sensitivity-conscious analyst.

Steps to follow

1. Update the data table in which the composite measure is calculated as new relevant data becomes available. Document the versions and annotate them so that other team members and your successors understand them.
2. Decide as a team which version will be communicated to the outside, and with what qualifications describing the remaining uncertainty in the data.

Back to the real data

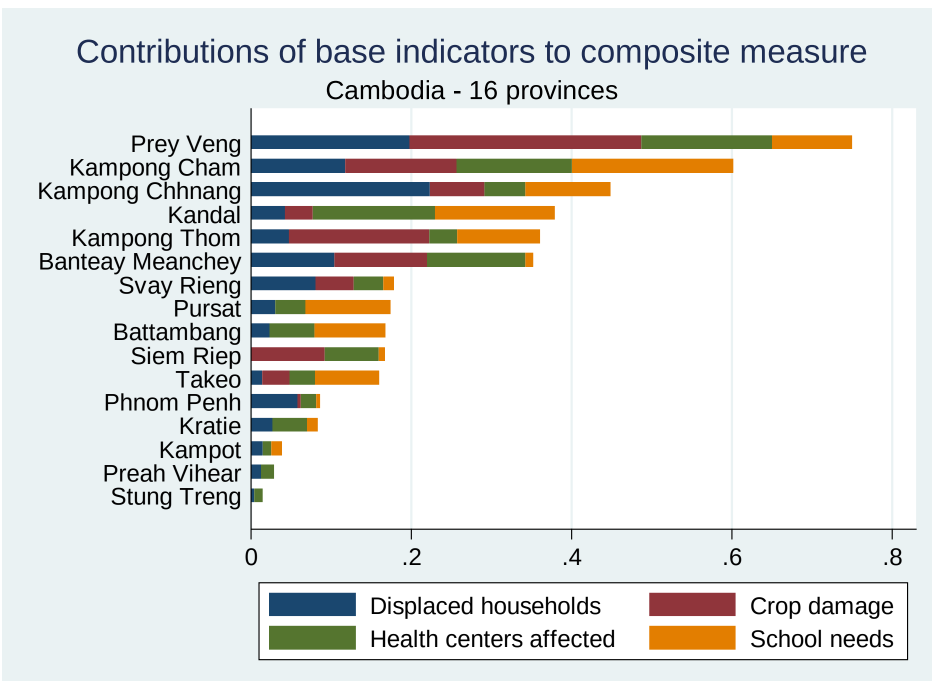
Considerations

Under this slightly bewildering title - as though we had strayed from the "real" -, the Handbook packs a smorgasbord of sundry next steps after the composite measure has been designed, computed and tested for robustness and sensitivity. Some of these belong rather in the section on presentation and visualization. Others, such as correlational analysis, supposedly were done earlier.

One specific recommendation that the Handbook makes is to *"decompose the composite"* measure and to *"document and explain the relative importance of the sub-components"*. The added value from this operation, however, is uncertain. We already know that under normalization by column sum and strictly additive aggregation, the contributions from the various indicators are proportionate to the weights. If the weights are equal, each indicator makes an equal contribution to the sum of composite measure values. This, we have seen (and applauded), is the case in the Cambodia design.

Another decomposition method in the Handbook, often seen also in other contexts, is by stacked bar chart. The segments of the bars represent the contributions of particular indicators to the composite value of the unit. This figure does so for the Cambodia data.

Figure 17: Bar graph of indicator contributions to composite



However, this is tantamount to a full dashboard visual, modified by normalizations and weights. For 16 provinces and 4 indicators in Cambodia, the chart carries 64 elements distinguished by length and color (this number includes the invisible zero elements). Other than for looking up a specific value of interest, this surely must overburden viewers. Moreover, while stacked bar charts are *formally* correct in depicting indicator contributions, they prove nothing about the *substantive* validity of the composite measure.

Steps to follow

If the previous steps were taken correctly, there is nothing to do here.

Links to other variables

Considerations

That composite measures should be linked to other variables, as the Handbook recommends, is sound advice. The linking will be within models that make sense in the - current or anticipated - disaster, needs, response and recovery contexts. Formally, a number of methods can be helpful - scatterplots, multiple bar charts (on suitably comparable scales), correlation matrices, regression models, overlaid or side-by-side maps. Obviously, the variable to which we link the composite must not be one of its own components.

The linking will in large part be opportunistic, depending on newly available data. Conversely, data on other variables that are equally desirable from a needs assessment viewpoint may not be available in time, or may be based on entities (e.g. camps instead of districts) that make them unsuitable for linkage. Proactively, at the time when the composite measure is being designed, the analyst may consider what other data could and should be collected one, two, three months later, of a kind that would make for powerful combinations with the impact measure. If the flood impact measure in Cambodia was based on reports as of November 2011, what was the child malnutrition rate in February 2012?

Particularly interested correlations may emerge, in the time after the first needs assessments, between the composite impact measures and subsequent response variables, and then again between those and recovery indicators.

One may speculate that links between composite measures derived from surveys and administrative reports on one hand and remote measurements of spatially defined entities, processed in GIS, on the other will have increasing explanatory and predictive power. For example, night light measurements from satellite might eventually be put into the public domain with shorter delays and with 1 km by 1 km resolution. Changes in luminosity could be assumed to be correlated with disaster impact and again with recovery, as observed for Rwanda 1992 - 1996 (Henderson, Storeygard et al. 2012: 1004).

Step to take

1. Cultivate a general open and creative attitude for the use, by yourselves and your partners, of the composite measures that you produced in combination with other information.

2. At the time of design and again analysis, think forward to later stages of the humanitarian action and propose types of data and data collection activities that could be fruitfully linked to the composite measure.

Presentation and visualization

Considerations

The communicative uses of composite measures differ according to measurement level that the assessment wishes to emphasize. This choice may be motivated also by the confidence that the analyst has in the validity of the measure and in the anticipated reactions by assessment users.

In its **ordinal** interpretation, the composite measure is simply a ranking device, suitable for sorting tables and designating the X most affected units. One can sort the original indicator table by the composite and still let the indicators speak for themselves, with the composite played down and explained in footnotes. Or one can give it more prominence, with color in the table and emphasis in the narrative.

The story is different when the composite is used at the **interval** level. Here the measure takes a leading role. We invoke it to create attention for outliers ("by far the worst impacted province"), groups distinctly set apart ("another group of several provinces, while significantly affected by the disaster, has escaped catastrophic damage"), or for a correlation with another important variable ("the four districts closest to the epicenter all suffered impacts three to four times larger than the rest of the surveyed areas").

Bar graphs and choropleth maps are particularly well suited to visualize the absolute (interval level) values of the composite measure. Maps emphasize the spatial clustering whereas bar charts let the viewer see proportions. Associations with an external variable can be expressed in multiple bar charts or in scatterplots, as mentioned earlier.

The choice of visuals should follow the narrative intent, all the more so in reports that tilt the balance between text and image to the latter.

Steps to follow

1. Determine the message that the composite measure is to support.
2. Anticipate what framing (introduction, placement, footnoting, connection to subsequent topics in the report) the measure will need in order for that message to be understood.
3. Then choose an appropriate tabular or graphic representation.

The decision to form a composite measure

After saying so much about the finer points of composite measures, it is important to once more apprehend whether a rapid assessment team should try to form any at all, and if so with what justification. We put forward some pointers in a goal - means - special conditions kind of exam schedule.

Purpose

Trivially, the composite measure should serve a useful purpose. We discussed several motivations earlier. One may be crude geographical targeting, which requires a rankable unified measure. A belief in the seriousness of cumulative effects of diverse disaster impacts could be another. Similarly, the need may be strong to simplify the jungle of raw indicators into something that the assessment consumers will appreciate within their limited attention spans.

It is, however, not sufficient to hint that the composite measure may have several benefits; one has to find clarity what he specifically expects the measure to achieve. What is the value added to the assessment findings?

Consistent meaning

The second key test question is: What does the measure really express? What does this number mean? If province A scores higher on it than province B, does it mean, for example, that

- the people in A are suffering more than those in B? (to show up in higher excess mortality in the next months if not addressed), or
- institutions and services in A have been disrupted more severely than in B (to show up in higher recovery costs in the next month and years), or
- the bases for further growth and development in A have been more badly damaged than in B (to show up in slower progress towards the Millennium Development Goals over the next decade)?

Once we have decided what we want to measure (e.g., the intensity of short-term unmet needs) and are in the process of designing the measure more specifically, we should save a reminder for further down the road: Is the measure truly transitive in the relations of the units for which we will have computed it? Consider the outcomes A = 3, B = 2, C = 1 for affected provinces and their composite scores. Suppose we are satisfied that indeed A > B and B > C in terms of the meaning of our measure (e.g. "unmet needs"). Given all the detailed knowledge from the base indicators, does the <u>same</u> interpretation hold for A > C? If we disagree that the unmet needs of A are greater than those of C, obviously our construct is not a consistent measure of "unmet needs" and needs to be revised or abandoned.

Feasibility

This aspect comes next. Two conditions must be met:

- The required data must be extractable from information on hand or collected within useful time and with defensible direct and opportunity costs.
- The assessment team must be able to analyze the data and to communicate the working of and findings from the composite measure.

The two parts are related. For example, missing values can be patched up to a degree if the team knows how to handle them. When the base indicators are numerous (we speculated earlier that 4 - 5 may be a critical threshold for teams limited to spreadsheet analysis), expertise in multivariate procedures, using statistical applications, becomes more important.

Other conditions: Acceptance and updating

Even if the first three challenges are all overcome, other obstacles may arise. Foremost among them is stakeholder acceptability.

- Do important stakeholders reject composite measures out of principle (because they want to preserve their freedom of interpretation looking at the full range of base indicators)?
- Do they object to particular design elements (disagreeing over the process of measurement, over weights or the treatment of units with missing values, etc.)?
- Do they object to a particular interpretation of the measure itself or of its association with some external piece of information?

If the team anticipates that its efforts to form composite measures will heighten dissension among the disaster responders, it may need to have second thoughts, more consultation or stronger political backing.

Some consideration - we cannot call it a condition - should be given to the process of updating, i.e. the likely changes of the information landscape as assessment and response proceed. Is it probable that missing values will be filled? Or are the indicators only of fleeting interest, soon to be replaced by other concerns? Can we anticipate the collection of other types of information with which the currently designed measure can be associated for some deeper insight into the post-disaster dynamic? Etc. This essentially is speculation on the conceptual longevity of the composite measure under construction.

In the real life of assessment teams, a discovery process is likely to take place in the opposite direction. Possessing some indicator data, the analytically interested member may assume that gaps will be filled, that reduced complexity will be appreciated by all, that the meaning of the measure can be guessed after its calculation, and finally that it will be seen as useful in different contexts. Such attitudes may be ambidextrous and creative, but they do not away with the necessity to decide the composite measures on solid grounds.

Outlook

The reader must have felt that this note breathes a skeptical spirit. Our skepticism focuses on the purpose as well as on the validity of composite measures formed and used in rapid needs assessments. What do such measures achieve that could not be delivered by other means? What is it that they really measure?

Encouraged by similar debates in international development, we propose alternatives that hold a middle ground between the all-encompassing composite measure and the dashboard of separate indicators. We believe that subindices can be formed of substantively grouped indicators. These constructs, suitably displayed, enable typologies of impacts and clusters of affected units by impact constellation. They avoid deciding the importance of all indicators at once while making the diversity of impacts visible.

This is not to deny the case for composite measures in needs assessments. Disaster impacts are correlated. They have cumulative effects on human suffering and on recovery burdens. When we express disaster effects quantitatively, it is legitimate to consider combined measures apt to capture the cumulative effects.

What matters, however, is that the design of such measures ought to be guided primarily by substantive considerations. Formal requirements - such as the trinity of normalization, weighting and aggregation - must be met, but they are less problematic. There are ideas, algorithms, standards that can be adapted from best-practice authorities such as the OECD Handbook. The formal limits - imposed by time pressure, skills, data quality - can, to a point, be stretched. The substantive challenges - the scarcity of validated, calibrated, transferrable impact measures - are harder to overcome. Technical sophistication does not resolve them.

Other observers may find reasons to be more optimistic. In emergencies, there is virtue in improvisation and in second-best solutions; this entitles us to work with ad-hoc constructs (as long as the uncertainty in the data is acknowledged, and an attempt made to define the relative importance of the indicators). In the longer run, needs assessment techniques move together with the humanitarian and disaster management communities. These partake in technological changes - survey data management, crowd sourcing, remote sensing, etc. - some of which will bring new opportunities to the designer of composite measures. In addition, research using the information that the new sources generate may improve on the validity of some measures while eliminating others.

There are encouraging developments to be noted also on the data collection front. Governments have the clout and increasingly the skills to collect, rapidly after disaster strikes, the kinds of quantitative data that provide the raw material for potential composite impact measures. In South and South-east Asia, for example, the governments of several countries - Bangladesh, Cambodia, the Philippines - have defined, and actively employed in recent disasters, tools for rapid data collection. While a great deal of methodological improvement is still needed, the signs that impacts can be reported with reasonable speed and coverage are growing stronger.

Just as there is now a "disaster risk equation" combining hazard, vulnerability and capacity, we may some day know a widely accepted "disaster impact equation" or a "humanitarian needs equation" - composite measures making rapid assessments more credible and more efficient.

Appendix

Robustness to changing weights

For composite measures, the effect of changing indicator weights can be rapidly simulated in a spreadsheet. However, the key device for this is the Excel function TABLE, the mechanics of which are not straightforward.

We demonstrate how it works, in sheet *"C_WeightsEmperim"* in the demonstration workbook, using the indicator set for the Cambodia floods assessment (see page 28). We are interested in change in the relative rankings of 16 provinces as the weight of a given indicator varies.

The worksheet has five areas:

- The indicator - composite measure table
- The weights table
- The simulation table
- A working copy of the simulation table for rearranging results
- A chart of the rank variability

For easier understanding, we set all the initial weights to 1 and adjust during aggregation, i.e. divide the sum of weighted indicators by the sum of weights. We vary, for one indicator at a time, the weight over a plausible range. Here we choose {0, 0.5, 1, 1.5, .. 4.5, 5} for discrete values in such a range.

Calculating the composite measure

In the indicator / composite measure table, insert columns to hold the reweighted indicators. Calculate their values by multiplying the original value with the weight noted in the weights table. If the weight cells are named (recommended for clarity), the formulas become intuitive, such as

*=RC[-1]*DisplacementWeight*

Name the sum of weights in R25C2 "Weights".

Calculate the composite measure by adding the reweighted indicator values and dividing this sum by the sum of weights, as in

=SUM(RC[-7],RC[-5],RC[-3],RC[-1]) / SUM(Weights)

Name the range of composite measure values R3C12:R18C12 (here it is called "Composite2"). In C13, calculate the ranks of the provinces by the composite measure:

=RANK(RC[-1],Composite2,1),

where the last parameter, set to 1, ensures that the lowest value is ranked as 1.

So far, nothing has been simulated. We can, of course, test for robustness manually, by changing the values in the Weights area, but it will be difficult to summarize the results from such an exercise.

Using the TABLE function

A more systematic simulation can be done using the Excel function TABLE. The Help page "Calculate multiple results by using a data table" is somewhat helpful for background, but the mechanics concerning this situation is better taught by way of example.

Create a simulation table on the lines of R29C1:R45C13 in our example. The weight values to be varied are entered in the top row, here in R29C3:R29C13. The references to the provinces must appear in numerical form, as the sequence in which values must be looked up in the range Composite2. We thus number them consecutively 1, 2, .. 16[18].

At the intersection of the weight values and the 1-16 sequence, in cell R29C2, we write the lookup formula for the ranks:

=RANK(INDEX(Composite2,R29C1,1),Composite2,1)

Also in R29C1, we write a starting value for the lookup sequence (1). The formula then works like this: *INDEX(Composite2,R29C1,1* looks up the value in the range Composite2 in the relative row as given in R29C1 and in its relative column 1 (trivially, because Composite2 has only one column). This value is then passed to the RANK function. In *RANK([composite value looked up by INDEX],Composite2,1)* "composite value" is the value for which to establish the rank among the elements of Composite2, and the last parameter, as mentioned, ensures that the ranks are number with the lowest value given 1.

Select the range that begins, in the upper left, with that formula cell and extend, in the lower right, to the last combination of lookup sequence and weight values. Here this is the area framed in a thick red border, R29C2:R45C13.

From the menu, call up Data - Data Tools - What-If Analysis - Data Table. For row cell input, select the weight cell for the particular indicator weight that you wish to vary. Here we chose the indicator "school needs" and its weight in cell R24C2. For column cell input, choose the lookup sequence cell, which we placed in R29C1. OK.

TABLE will now vary the lookup sequence number from 1 to 16, and the value of the selected indictor value from 0 to 5, and will compute and write into the table the rank for all combinations. In the final display, the values displayed in R24C2, R29C1 and R29C2 are set back to their starting values.

---

[18] Note that in the Cambodia case, this is slightly different from the original sort order shown in the indicator table. Here, the provinces initially numbered 9 and 10 have been deleted because of missing values. The original sort order is simply kept for historical reasons.

If done correctly, the computed ranks should appear as in R30C3:R45C13. Use conditional formatting with a convenient color ramp for high to low ranks. This creates a heat map for easy visualization. Note that at this stage you can manually change the weights for indicators other than the one referenced in the formula in R29C2. For example, you can give a weight of 3 for displaced households. The simulation table will automatically recalculate. Whether this makes sense is a substantive question.

Make a working copy somewhere in the sheet (in the Table area, the formulas will automatically be replaced by their values). Edit as convenient, such as by sorting on the rank values for the original weight = 1 (take care to select the entire area with the green border!) and by reformatting such as in Table 4 on page 31.

In order to simulate rank changes for variations in the weight of a different indicator, return to the simulation table, select again the red-bordered area and define row and column input cells in Data - Data Tools - What-If Analysis - Data Table[19].

Alternative aggregation functions

The OECD Handbook states that [not only weighting methods but] *"aggregation methods also vary. While the linear aggregation method is useful when all individual indicators have the same measurement unit, provided that some mathematical properties are respected. Geometric aggregations are better suited if the modeller wants some degree of non compensability between individual indicators or dimensions. Furthermore, linear aggregations reward base-indicators proportionally to the weights, while geometric aggregations reward those countries with higher scores"* (op.cit., 32-33).

However, as an alternative to linear aggregation (which amounts to a weighted arithmetic mean), neither the geometric mean nor the non-compensatory aggregations that the authors recommend are suitable in our situation:

- The geometric mean, while easy to calculate as $(x_1 * x_2 * .. * x_n)^{(1/n)}$, returns zero whenever at least one of the base indicator values is zero. In situations where we are bound to measure "welfare bads", it is likely that for some units and indicators, zeros occur. Returning zero as the aggregate value for these units is likely to be seriously misleading.

- Non-compensatory aggregation methods are analytically difficult and not all are feasible with the standard spreadsheet features.

Instead we propose, and have experimentally used on the Global Focus Model data, a class of aggregation functions that were used also in the Human Development Index calculations. While they require an initial effort to understand how and why they work, they are straightforward to compute with Excel formulas.

---

[19] Global replacement with Ctrl + H and modification in the formula bar with array formula entry do not work with TABLE.

The HDI modelers, anxious to integrate Amartya Sen's perspective on capabilities and deprivations, wanted an aggregation function for their poverty index with certain properties. It should return an index value that would tend towards the maximum of the (rescaled) indicator values for a given country. This was how they constructed the "Human Poverty Index" (Anand and Sen 1994; UNDP 1998), which in recent years was replaced by the "Multidimensional Poverty Index".

So called L-p-norms (Wikipedia 2012a) define a class of functions that meet the poverty index requirement. The same type of functions also serves the purpose of (relatively) adjusting the risk index value upward if the indicator values for a unit are relatively similar, and downwards if they are dissimilar. Such an adjustment might reflect, as we argue in the main body of this report, the theory that multiple disadvantages have a disproportionate effect on the risk of humanitarian emergencies.

It is best to demonstrate the mechanics with a graphic example. In the graph below, the index y is a function of two indicators x_1 and x_2. These range from 0 to 10, as in the GFM. To show the shape of the functions in two-dimensional space, we fix one of the indictors to an arbitrary value. Here we set $x_2 = max[0, 10] = 10$.

We are now trying out different values of the parameter a in the Lp-norms for x_1, x_2, which we give the shape

$$y = \underbrace{(x_1^a + x_2^a)^{(1/a)}}_{\textbf{\textit{Lp-norm}}} * \underbrace{(max(x_1, x_2) / (2 * max(x_1, x_2)^a)^{(1/a)})}_{\textbf{\textit{Rescaling factor}}} ,$$

which simplifies to

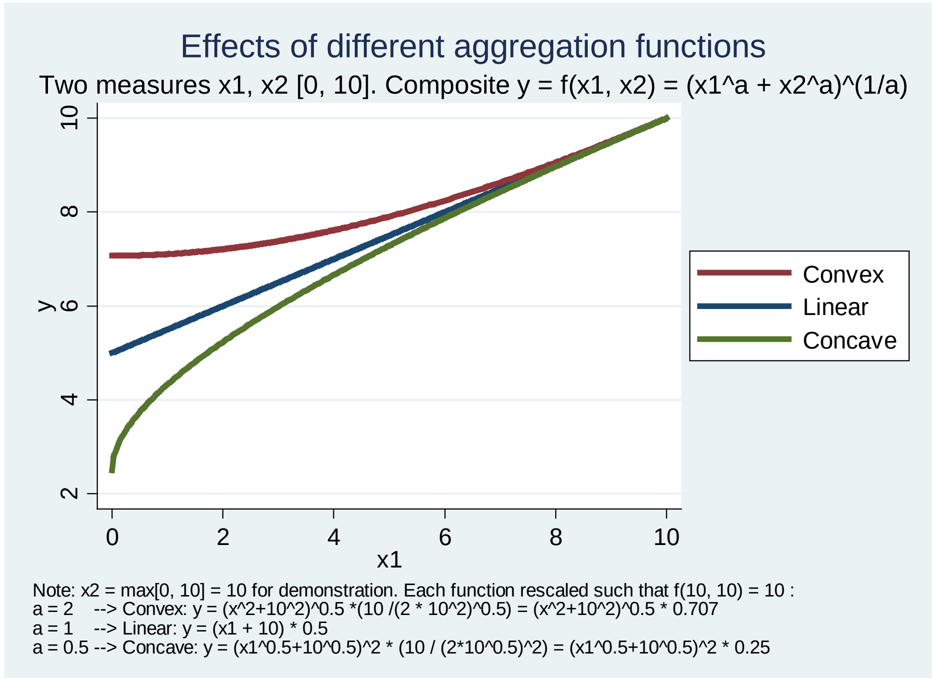
$$y = (x_1^a + x_2^a)^{(1/a)} * 2^{(-1/a)}$$

In the case of x_1 $[0, 10]$ and $x_2 = 10$, we get

$$y = (x_1^a + 10^a)^{(1/a)} * 2^{(-1/a)}$$

The diagram shows the curves for a = 1 (the basic linear case), 2 (similar to the one used in the HDI poverty index), and 0.5. We use this last one to reflect cumulative disadvantage. The rescaling factor ensures that y remains bounded in [0, 10].

Figure 18: Examples of Lp-norm aggregation functions



For the experiment with the Global Focus Model data, we chose an even more concave function, with a = 1/3. The implementation in Excel is shown in this screen shot.

Figure 19: Global Focus model country risk with different aggregation functions

R2C22 = (RC7^(1/3) + RC12^(1/3) +RC16^(1/3))^3 / 27

	1	7	12	16	17	22	23	24
1	Country	Hazard	Vulnerability	Capacity	Risk	Risk (altnative aggregation)	RiskRank	RiskAltRank
2	Congo DR	6.65	9.11	7.47	7.74	7.70	147	147
3	Sudan	6.76	8.52	7.68	7.65	7.63	146	146
106	Marshall Islands	0.50	4.26	6.73	3.83	2.95	43	25
107	Tuvalu	0.50	4.91	6.08	3.83	2.98	42	28
123	Oman	2.41	2.91	4.42	3.25	3.17	26	36
127	Argentina	3.02	2.66	3.50	3.06	3.05	22	31

with examples of counties whose risk ranking remained unaffected (Congo, Sudan), was lowered (Marshall Islands, Tuvalu), or was bumped up (Oman, Argentina). The Excel formula for the alternative risk score:

$$= (RC7^{\wedge}(1/3) + RC12^{\wedge}(1/3) + RC16^{\wedge}(1/3))^{\wedge}3 / 27$$

is composed, as before, of the Lp-norm and the re-scaling factor (* 1/27). Note that *all* the alternative aggregation-calculated risk values are lower than the original GFM values; this is so because of concave form and of the re-scaling factor such that f(10, 10, 10) = 10. The point of the experiment is to demonstrate changes in the ranking.

The results for the entire groups of countries were summarized in the main body.

Excel demonstration workbook

Currently, the demo workbook is named *"120503AB_CompositeMeasures_Demo.xlsx"*.

It has the following worksheets:

Table 8: Worksheets in the Excel demo workbook

Generalities	
BasicSetup	Basic operations in computing a composite measure
Global Focus Model	
GFM_OriginalTable	Original data table
GFM_Variables	Variables used
GFM_Data	Same data, formatted as normal Excel data table
GFM_AggregAlternat	Rank changes when alternative aggregation formula is used [see page 58 of the note]

[cont. next page]

Cambodia floods	
C_OriginalTable	Original data table
C_WorkingCopy	Same data, with missing values imputed to zero plus some variables added from other tables
C_Variables	Variables of the normally formatted Excel table
C_Data	Same data, formatted as normal Excel data table
C_WeightExperim	Demonstration of a weight-robustness test, using the TABLE function [see page 56 of the note]
C_WeightsFormatted	Convenience copy of the results formatted for the report
C_SubindicesData	Formation of two subindices [see pages 7 and 30 of the note]
C_SubindicesSummary	Pivot table to summarize impact typology in table format
C_LSmethodProcess	Process of calculating ranks by the lexicographic semiorder method [see page 46 of the note]
C_LSmethodCompare	Results of calculating ranks by the lexicographic semiorder method; comparison with other methods
C_MissingImputed	Demonstration of a median-(instead of zero-)based imputation of missing values in the indicators
Auxiliary table	
ListOfNamedRanges	Convenience table of all named ranges used in this workbook

References

Acaps. (2011a)."Quick Impact Analysis - Earthquakes [Version 14 October 2011]." Geneva, Acaps. Retrieved 12 April 2012, from http://www.acaps.org/img/documents/earthquakes-quick-impact-analysis-earthquakes.pdf.

Acaps. (2011b)."Quick Impact Analysis - FLOODS [Draft 14 October 2011]." Geneva, Acaps. Retrieved 12 April 2012, from http://www.acaps.org/img/documents/floods-quick-impact-analysis-floods.pdf.

Anand, S. and A. Sen (1994). "Human Development Index: methodology and measurement." Human Development Occasional Papers (1992-2007).

Benini, A. A., C. Conley, et al. (2009). "Survivor Needs or Logistical Convenience? – Factors shaping decisions to deliver relief to earthquake-affected communities, Pakistan 2005-06." Disasters **33**(1): 110-131.

Bobko, P., P. L. Roth, et al. (2007). "The usefulness of unit weights in creating composite scores." Organizational Research Methods **10**(4): 689.

Chen, X., H. Hong, et al. (2011). "Nonlinear models of measurement errors." Journal of Economic Literature **49**(4): 901-937.

Cherchye, L., W. Moesen, et al. (2007). "An introduction to ‘benefit of the doubt’composite indicators." Social Indicators Research **82**(1): 111-145.

Comfort, L. K., N. Oh, et al. (2009). "The Dynamics of Disaster Recovery: Resilience and Entropy in Hurricane Response Systems 2005–2008." Public Organization Review **9**(4): 309-323.

Cooper, W. W., L. M. Seiford, et al. (2007). Data envelopment analysis : a comprehensive text with models, applications, references and DEA-solver software. New York, Springer.

Ferreira, F. H. G. and M. A. Lugo (2012). Multidimensional poverty analysis: Looking for a middle ground [Policy Research Working Paper 5964]. Washington DC, The World Bank.

Finch, C., C. T. Emrich, et al. (2010). "Disaster disparities and differential recovery in New Orleans." Population & Environment **31**(4): 179-202.

Garfield, R., C. Blake, et al. (2011). Common Needs Assessments and humanitarian action [HPN Network Paper No. 69, January 2011]. London, Overseas Development Institute (ODI).

Gorst-Rasmussen, A. (2012). "tt: Treelet transform with Stata." The Stata Journal **12**(1): 130-146.

Henderson, J. V., A. Storeygard, et al. (2012). "Measuring economic growth from outer space." American Economic Review **102**(2): 994-1028.

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

Kukla-Acevedo, S., M. E. Streams, et al. (2012). "Can a Single Performance Metric Do It All? A Case Study in Education Accountability." The American Review of Public Administration **42**(3): 303-319.

Lansdowne, Z. F. (1996). "Ordinal ranking methods for multicriterion decision making." Naval Research Logistics (NRL) **43**(5): 613-627.

Luce, R. D. (1956). "Semiorders and a theory of utility discrimination." Econometrica, Journal of the Econometric Society: 178-191.

Manzini, P. and M. Mariotti (2012). "Choice by lexicographic semiorders." Theoretical Economics **7**(1): 1-23.

Nardo, M., M. Saisana, et al. (2008). Handbook on Constructing Composite Indicators: Methodology and user guide. Paris and Ispra, OECD and the Joint Research Centre (JRC) of the European Commission.

NodeXL Team. (2012)."NodeXL: Network Overview, Discovery and Exploration for Excel." Social Media Research Foundation and Microsoft Corporation. Retrieved 10 April 2012, from http://nodexl.codeplex.com/.

Offer, A. (2000). Economic Welfare Measurements and Human Well-Being [Discussion Papers in Economic and Social History. Number 34, January 2000]. Oxford, Nuffield College, Oxford University.

Ravallion, M. (2010). Mashup Indices of Development [Policy Research working paper # 5432]. Washington DC, The World Bank.

Sauder, M. and W. N. Espeland (2009). "The discipline of rankings: Tight coupling and organizational change." American Sociological Review **74**(1): 63-82.

Tucker, L. R. and R. C. MacCallum. (1997, 14 April 2012)."Exploratory factor analysis." Internet publication. from http://www.unc.edu/~rcm/book/ch1.pdf.

Tversky, A. (1969). "Intransitivity of preferences." Psychological Review **76**(1): 31-48.

UNDP (1998). Human Development Report 1997. Annex: Technical and Statistical Notes. New York and Oxford, Oxford University Press**:** 240 pp.
UNOCHA (2011). Kingdom of Cambodia: Evaluation of Post-Flood Needs Assessment Data. Geneva, United Nations Office for the Coordination of Humanitarian Affairs.
UNOCHA and Maplecroft (2011). 2012 Global Focus Model [27 December 2011]. Geneva, Switzerland and Bath, United Kingdom, United Nations Office for the Coordination of Humanitarian Affairs and Maplecroft.
Vichi, M. and H. A. L. Kiers (2001). "Factorial k-means analysis for two-way data." Computational Statistics & Data Analysis **37**(1): 49-64.
WHO and King's College London (2011). The Humanitarian Emergency Settings Perceived Needs Scale (HESPER): Manual with Scale. Geneva, World Health Organization.
Wikipedia. (2009)."Biplot." Retrieved 3 April 2012, from http://en.wikipedia.org/wiki/Biplot.
Wikipedia. (2011)."Borda count." Retrieved 28 March 2011, from http://en.wikipedia.org/wiki/Borda_count.
Wikipedia. (2012a)."Lp space." Retrieved 3 April 2012, from http://en.wikipedia.org/wiki/Lp_space.
Wikipedia. (2012b)."Normal score." Retrieved 13 April 2012, from http://en.wikipedia.org/wiki/Normal_score.
Wikipedia. (2012c)."Semiorder." Retrieved 18 April 2012, from http://en.wikipedia.org/wiki/Semiorder.
Yoon, K. P. and C.-L. Hwang (1995). Multiple Attribute Decision Making. An Introduction [Quantitative Applications in the Social Sciences #104]. Thousand Oaks, Sage University Paper.

18 April 2012 / Revised 17 May 2012 / AB

For related resources, please visit http://www.acaps.org/en/resources.