Aldo Benini
A note for Okular Analytics

Priorities and preferences in humanitarian needs assessments -

Their measurement by rating and ranking methods

23 January 2019

Contents

Tables

Figures

Acknowledgment

The section on Copeland's Method was inspired by an Excel workbook accompanying

> *"Saisana, Michaela. (2016). Step 6: Aggregation rules: Non-compensatory approaches. COIN 2016 - 14th JRC Annual Training on Composite Indicators & Scoreboards, 26-28/09/2016. Ispra (IT), The European Commission's science and knowledge service and Joint Research Centre."*

For our companion workbook we have significantly modified worksheet architecture and cell functions. The contribution of Professor Saisana, who was the first to implement Copeland's Method in an adaptable Excel format, is gratefully acknowledged.

Without the help of Attilio Benini, Zürich, we would not have been able to reshape published R code of the Plackett-Luce model in a way that makes it easily adaptable by interested readers.

Summary

Humanitarian needs assessments seek to establish priorities for action and preferences of those whom the action will affect. Assessment teams collect data meant to capture the severity of problems and the situational aspects that matter in the evaluation of response options. When samples of stakeholders, such as key informants or relief workers, state priorities or preferences, the information chiefly produces ratings or rankings. Such data are ordinal. In summarizing them, we avoid operations that require metric variables (such as the arithmetic mean); yet we seek methods that produce aggregates with metric qualities (such as Borda counts). These have a higher information value and offer greater flexibility for subsequent analyses.

Table 1: An example of a rating-based measure

Key informants in 1,807 camp points in Rohingya Bangladesh rated the severity of unmet needs in March 2018 (NPM Round 9). The five-level scale goes from "not severe" (1) to "extremely severe" (5). The median rating is a first, if very coarse, measure of comparative severity.

Need	Severity rating					Total	Median rating
	1	2	3	4	5		
Cash	0.2	1.8	5.0	30.0	63.0	100.0	5
Firewood	1.2	0.7	2.8	37.5	57.9	100.0	5
Water	1.8	5.1	15.5	34.1	43.4	100.0	4
Jobs	0.6	5.1	14.8	43.5	36.2	100.0	4
Shelter	1.3	5.7	23.6	42.7	26.7	100.0	4
Food	0.8	5.7	24.0	48.4	21.1	100.0	4
Security	1.7	10.5	28.8	41.3	17.7	100.0	4
Education	3.5	9.8	34.4	37.0	15.3	100.0	4
WASH	2.4	15.5	40.6	31.6	9.9	100.0	3
Health	1.9	10.6	42.3	35.9	9.2	100.0	3
PSS	6.1	30.4	34.8	20.9	7.9	100.0	3
Hygiene	4.2	30.7	46.0	13.0	6.1	100.0	3
Transport	7.1	31.0	43.0	15.1	3.8	100.0	3
Training	5.6	32.5	44.4	14.2	3.2	100.0	3
Utensils	12.8	22.5	36.3	25.2	3.2	100.0	3
Registration	9.2	27.4	36.8	24.8	2.0	100.0	3
Clothing	6.6	30.3	44.4	18.0	0.7	100.0	3
All ratings	*3.9*	*16.2*	*30.4*	*30.2*	*19.3*	*100.0*	

Note: Population-weighted. Row-wise percentages. Sorted descendingly on percentages of level 5.

This note is about the aggregation of rating and ranking data and about the interpretation of the resulting measures of priority and preference. Two distinctions are fundamental. The first concerns statistical independence. Ratings (such as of the severity of unmet needs in various sectors), though cognitively interdependent among items, statistically are independent. Rankings reflect an order among items and are dependent in both respects. For the analysis, both types have pros and cons; they contribute the most when used in

tandem. This can happen during data collection – by asking both rating and ranking questions – and/or in the analysis – such as by ranking aggregate measures from ratings.

Second, certain measures characterize cases (units like geographical areas, communities, camps, eventually households and individuals). Others express relationships with items (formally: options, aspects, attributes, indicators; substantively: sectors, agencies, problems, etc.). Methods suitable to compare Case A vs. Case B as well as Item X vs. Y are few; most of those based on ordinal data work only in one or the other way, not both.

The note is organized around those two distinctions. For measures aimed at comparing cases (e.g., how serious the WASH needs are in a number of IDP camps), we discuss two methods (#1 and 2), both of them based on ratings. To compare items (e.g., the severity of unmet needs between sectors for the ensemble of assessed communities), we discuss one method for ratings and two for rankings (#3 – 5). There are a great many more methods and approaches out there in the vast field of statistics. We focus on those that users can carry out with the tools of Excel, the workhorse of humanitarian data analysis. Only for #5 do we select a program from a statistical application; the benefits justify the extra effort.

1. **Rank-statistics of ratings to compare cases:** When cases are rated on a number of aspects (e.g. the levels of unmet needs), and the raters use the same scale for all aspects, the *case-wise medians* are a suitable measure of a common underlying construct (in this case, the severity of deprivation). The measure, by the definition of a median, is *ordinal*. By transforming the ratings appropriately (e.g., to ridits), a *metric* measure can be obtained. However, since infinitely many transformations are possible, the arbitrariness may deter users from taken this avenue. If the major interest is to test for significant differences between groups (e.g. residents vs. IDPs), transformations are not needed; a test exists for ordinal variables.
2. **Copeland's method** works for ratings as well as for combinations of ratings and other data types. This flexible method admits items that use different rating scales. It incorporates binary, interval or ratio-level indicators. The indicators may have different weights. Copeland produces a score (which can be interpreted as an interval-level measure) and, on the basis of this, a ranking of cases, i.e. an *ordinal* measure. It is unknown whether a test for differences between groups exists.
3. **Rank-statistics of ratings to compare indicators:** When ratings on several indicators follow the same scale, their *indicator-wise medians* can be compared. Again, this is ordinal. A convenient (in Excel) so-called "sign test" evaluates the difference in the distributions of the ratings between a pair of indicators.
4. For dealing with **ranked preferences** and priorities, the *Borda count*, because of its simplicity, is an almost indispensable measure. Needs assessment elicit only a partial variant, usually the first, second and third choices. This restricts the interpretation to differences in the intensities of preferences; ratios of the kind "the need for A is x-times stronger than that for B" are not supported. Thus, Borda supplies an *interval-level* measure. Group-wise differences within a given item can be tested for statistical significance, but the validity of the test is unclear.
5. **The Plackett-Luce model of rankings** overcomes this barrier. It produces a true *ratio-level* measure of the strength of preferences from full or partial choices. Tests

for differences between items follow from confidence intervals; they can be extended for groups within a given item. Using data from the Rohingya refugee camps in Bangladesh, we present estimates for priorities among 13 areas of need. We also provide code for similar calculations in the statistical application R.

The sections discussing methods #1, 3 and 4 are elaborations of practices that are fairly commonplace among humanitarian analysts. Here railguards are needed chiefly against the temptation of treating ordinal data as metric (the geometric mean of ratings is a frequently committed abuse) and for a correct understanding of the Borda count. By contrast, Copeland's method (#2) and the Plackett-Luce model (#5) may be novelties for many in this community. Both make additional demands, in terms of concepts and programming, but the benefits – acceptance of various data types by the first, a true ratio-level measure of priority by the second – will be increasingly understood and sought after.

In-between, we touch upon some of the pros and cons of population-weighting and of rating problems at the subsector level. Population-weighting yields to the natural impulse to give greater influence to larger communities such as cities, and less to villages, hamlets and nearly depopulated places. However, assessment data suffer from measurement error; if bias is correlated with population size, estimates may be further distorted. Therefore, if time permits, one should work out key statistics both weighted and unweighted; if the differences are important, they must be reported and, as far as possible, explained.

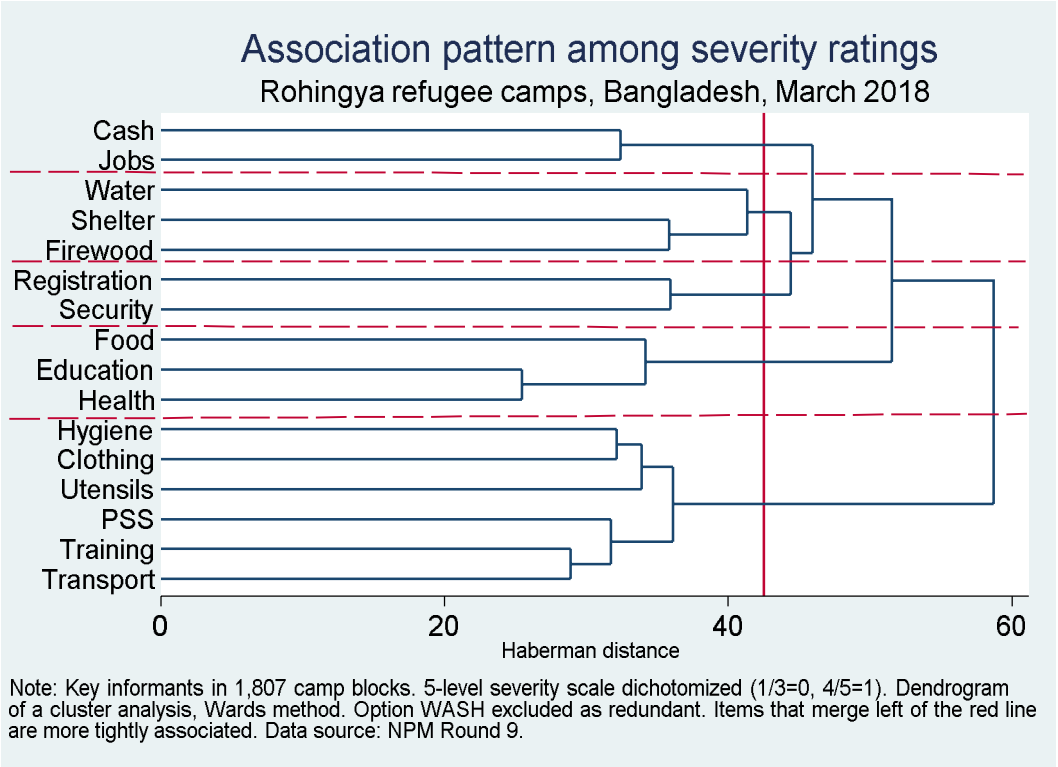
We are skeptical about tying problems firmly to particular subsectors within a larger sector of interest. The provision, use and drainage of water, for example, shades into all departments of the WASH sector. Before we build measures of priority and preference from items that were artificially sub-divided, we ought to dive into the deep structure of indicators. This may call for methods of multi-variate analysis that are less commonly practiced, such as cluster and principal component analyses of ordinal variables.

This note is detail-oriented, down to the arcana of the five methods discussed. Over-arching recommendations are hard to make. Philosophically, methods that bundle areas of immediate physical need, protection concerns as well as needs with longer-term consequences (notably education) into comprehensive measures of priority should be questioned on reliability and validity grounds. Separate measures for each of those complexes may help to avoid cognitive overload during data collection (key informants given too many items to consider at once). They may reveal associations among needs for individual survival, dignity and collective agency that are instructive to multi-sectoral response plans with intersecting time horizons and agency networks.

At the technical level, there is a similarly fruitful tension between poles of sophistication. At one pole, we have methods that are simple, fast and respectful of what non-technical assessment consumers understand. Comparisons on the basis of median ratings, or of proportions of affected populations that key informants placed at certain levels of a severity scale, pass this test. On the other extreme, complex methods promise to exploit more data, deliver stronger measures or reveal patterns that simple summaries obscure. Copeland and

Plackett-Luce belong here. The latter has potential to deliver strong visual profiles of preferences that vary by group and circumstance.

Figure 1: An X-ray image of severity ratings



Note: Key informants in 1,807 camp blocks. 5-level severity scale dichotomized (1/3=0, 4/5=1). Dendrogram of a cluster analysis, Wards method. Option WASH excluded as redundant. Items that merge left of the red line are more tightly associated. Data source: NPM Round 9.

Five clusters of associated needs can be discerned. Unsurprisingly, cash and jobs are closely related, but jobs and training are not. Food, education and health care form a substantively more heterogeneous cluster; whether it reflects primary concerns of women might be an interesting question for further research.

Between those poles, there are tools and methods of defensible difficulty. We note two:

- For rating data, consider appropriate transformations to **metric constructs**. They replace habits that analysts desperate for quick summaries have practices with inappropriate statistics, out of a legitimate need. The ridit is a good starting point. This note merely touches upon the costs and benefits of transformations; more study, and ultimately empirical validation, are required.
- With rankings of needs areas or problems, the **Borda count** offers relative simplicity and the ability to identify a group of high priorities. “Relative” and “group” are important qualifications. The analyst must interpret the findings within the limits of an interval-level measure, but must not deter consumers with technicalities of measurement levels. From our comparison with the stronger Plackett-Luce, we are confident that Borda validly identifies a group of high-priority needs even if the two methods differ about the order within this group. In other words, until Plackett-Luce becomes more popular, use the Borda count.

This recommendation may be expanded into a final one for this note. Needs assessments must be rapid, transparent, and balancing cost vs. quality. This puts a premium on simplicity. The statistics, for the most part, are descriptive. The composite measures, although broken down by group and geographic entities, are unidimensional. Most of the samples are purposive, defying estimates of uncertainty. What is missing is X-ray machines for a deeper look into the data. Around the needs assessment community, there is a lot of methodological powering-up going on. Some of it, notably from the GIS revolution, is fully incorporated. Yet, the bridges can be made stronger. Reinforcement seems to be overdue in multi-variate methods, some of which involve probabilistic models. The choice of some of these in this note is almost accidental, but the direction is likely to align with wider developments that now present possibilities and soon, perhaps, requirements.

Introduction

What this is about

The focus of this note is on the analysis of measures of preference and priority in humanitarian needs assessment. Specifically, we look at data generated at an ordinal measurement level, primarily ratings and rankings. Other ordinal measures, notably count and continuous variables estimated in intervals, are not considered unless they are used as ratings or rankings within composite measures. Neither are we reprising earlier work on subjective measures relevant particularly to the data collection stage (Benini 2018).

Three major concerns structure the note. First, clarity is needed as to whether a given method produces a measure about cases, or about variables, or about both. Second, methods are particularly prized that combine ordinal data in measures of a higher level, i.e. interval or ratio-level. Such levels create stronger information value for the needs assessment than comparisons based merely on ratings and rankings can achieve. Third, different methods produce measures that can, or cannot, be broken down by groups (e.g., types of affected persons, regions). Are there test for significant differences for probability samples?

Ordinal measures in needs assessments

The motivation for clarifying data constellations and analysis options arises from the generic context of needs assessments. The needs of persons affected by crises and conflicts are multiple. They are commonly mapped to different nodes of humanitarian action known as "sectors" and "clusters". For these, individually and in concert, a key decision problem is to select from a choice menu of target groups, activities and resource types. Rationally, allocations express preferences based on reported deprivations (priority needs) and on the ability to fill gaps in provision (access and resources). Needs assessments establish priorities whereas response plans seek viable compromises between priorities and the means to address them.

However, available measures of priority differ widely by their levels of information. "Hard measures" that are intuitively credible and make for easy comparison are limited to population figures, incidence and prevalence counts as well as proportions based on them. Other *ratio-level* measures in the deprivation context include shortfalls in household income and in caloric requirements. Rarely, an *interval-level* measure may be available, such as historic means of outside temperatures during the cold season that will put at risk displaced persons in unheated accommodations.

Often, sectoral priorities, the relative importance of problems within a given sector, or preferences for particular response options are expressed in "soft measures". In particular, persons speaking for local groups of affected persons may be asked to rate or rank needs and problems. As noted, this type of information produces *ordinal* variables. In yet other situations, the basic measure of need will only be *nominal*, stating a distinct status (e.g., resident, IDP, refugee) or the binary absence or presence of a condition (there is a school in the locality, or not). Conflations of scales occur; for example, "school is nearby / is somewhat far away / is very far away / no school" conflates a nominal and an ordinal measure that will then conveniently be treated as ordinal.

There is a common belief that nominal, interval and ratio-level measures are easy to aggregate, "aggregation" here being loosely understood both as creating composite measures at the unit level and as averaging at higher group or geographical levels. By contrast, ordinal measures supposedly resist aggregation or produce measures of low discrimination only (such as when the medians of ratings across several indicators are the same for the vast majority of cases). This belief is correct when issues of reliability, normalization and weighting are minor. In such comfortable situations, ordinal data are indeed less tractable than the other major types. In needs assessment data, such comfort is the exception rather than the rule.

The advantages of interval and ratio-level measures are strong, given the types of statistical operations that they admit (differences, means, and in the latter case ratios). Understandably, then, there has been a long-standing endeavor, overflowing into a large literature, in the social, cognitive and decision sciences to take ordinal measures to a higher level. This can take two forms. In the first, a composite measure, combining several indicators (not necessarily all of them ordinal), yields a score for each case (or at least for each with complete values).

The second obtains when several ordinal measures conceptually relate to a common underlying concept, such as deprivation. The corresponding construct may then be thought of as a continuum from none to lethal. Ideally, it can be expressed as a ratio scale, with a meaningful zero point for "normal". In the absence of a zero point, one may still be able to defend an interval-level interpretation. Of interest are the locations that the originally ordinal indicators are assigned on the scale, one value per indicator *for the entire sample* (or one for each sufficiently large sub-sample if the construct can be calculated separately). Individual unit-level values usually cannot be obtained.

Our objective is to discuss a small number of methods that produce measures of priority or preference from ordinal data, either case-wise scores or indicator-wise values on a common construct. We proceed as follows. At first, we recall basic differences, strengths and limitations in ratings and rankings. Then we present a simple "decision tree" branching into five appropriate methods, depending on case vs. indicator focus and data type. We discuss each of those, either with a worked example or referring to published notes that present examples in sufficient detail.

Ratings and rankings

Definitions

As noted, ratings and rankings account for the majority of ordinal data generated in the course of needs assessments. A rating conveys the level of an item on an ordered set of elements (such as a scale from "entirely disagree" to "entirely agree"). The meanings of the elements hold beyond the items at hand (they are inter-subjective and apply to many more situations). However, the evaluation that the rater performs in order to assign an item to the level of his/her choice is, to various degrees, subjective. Formally, the ratings given different items on the same scale are independent.

By contrast, a ranking conveys the place of an element in an ordered set of items. The order may result from individual preferences of the person evaluating the listed items or from the distribution of an attribute on which most or all observers agree, such as population size as the basis for ranking cities. The ranks are mutually dependent because the ranking forces an order (even when equal ranks are allowed). Some or all of the ranks change when items are added to, or removed from, the set.

Strengths and weaknesses

For a more detailed discussion, the reader is referred to the resource *"Severity and priority - Their measurement in rapid needs assessments"* (Benini 2013b:6-8, 10, 18-20)[1]. The following table summarizes some of the known strengths and weaknesses of ratings and rankings (ibd.: 8) (largely based on: Roszkowski and Spreat 2012).

Table 2: Strengths and weaknesses of rating and ranking

Rating		Ranking	
Pro	Contra	Pro	Contra
Equivalence between items is validly expressed by equal levels on the scale.	Low differentiation (= overuse of the same score across items) due to social desirability or extreme response bias	Forces differentiation	Differences between items may be artificial if subjects view them as equivalent.
Degrees of difference can be expressed on the scale			Rank differences do not express degrees of difference.
Less time-consuming in interviews.	Attention often superficial, unfocused.	Respondents pay more attention.	More time-consuming in interviews.
			First and last items on a list tend to be over-ranked, middle items under-ranked.
	List dependency: Although each item is formally independent, the number of items influences how each is rated (learning effects during the interview).		List dependency: Whether subjects choose X > Y or Y > X, depends on how many other items are to be ranked. Lower ranks very unreliable. If many items are ranked, all unreliable.

[1] Available at http://aldo-benini.org/Level2/HumanitData/Benini_forACAPS_SeverityAndPriority_2013.pdf .

Item ratings are statistically independent.			Statistical dependence of ranks makes certain analyses problematic.
Both subjects and items can be scored (e.g. by medians or frequencies of values in ranges of interest)	Ordinal data: legitimate stats limited to frequencies, medians, minima and maxima	If subjects understand limited choice situation, Borda counts on interval level (ratio only if meaningful zero point)	No total score for subjects (since rank sum equal for all)

Rankings and ratings may be elicited within the same data collection encounter, such as in the interview with a key informant. They supplement each other but can substitute for each other in small measure only. The purposes and analytic potentials are different, as the next section will elaborate.

Moreover, when the same item – e.g., the degree of deprivation in a given needs area – is the subject of both ratings (on some scale) and rankings (compared to other items), the two measures are likely to be correlated, but the correlation may be weak. This was demonstrated for sectoral severity (rated) vs. priority (ranked), empirically and with simulated data[2]. The absence of strong correlations may make these measures seem disorienting and ultimately useless for purposes of response planning. It is the responsibility of the analyst to look into the joint distributions of items that were both rated and ranked and to reconcile importance and preference through careful interpretation[3].

The response – Perceptual and decisional aspects

Before analysis, of course, the stages of design and collection take place. Assessment design and interviewer behavior determine, to a significant degree, two aspects of the response:

- The respondents' clear, fuzzy, or absent understanding of the items (perceptual aspects).
- The respondents' uncertainty in relating a particular item to the choice set – the levels of the scale for ratings, the other items in the list for rankings (decisional aspects).

How to take good care of the perceptual aspects is the grist of survey methodology books. The first consideration for the decisional ones concerns the number of levels (rating scale)

[2] See Benini (op.cit.). In a large needs assessment in northern Syria, the (population-weighted) correlations between the severity and priority scores within each of five sectors varied widely, from -0.35 for shelter to +0.81 for WASH. In the case of shelter, measurement error can be demonstrated (pages 2 - 21). A simulation with seven sectors found correlations ranging from +0.22 to +0.67 under the assumption of no measurement error. With increasing error, they attenuate progressively (pages 36 – 41).

[3] "We recommend a combined rating–ranking approach to determine the preference order for all [items] in a set, including those considered to be of moderate importance" (van Herk and van de Velden 2007:1096).

and of items (to be ranked) – if it is too low, the respondent is forced into unwilling or senseless choices; if too high, levels and items exceed short-term memory or demotivate by the tedium of irrelevant ones.

We leave this short and superficial section in order to leap to the topic of primary interest, the analysis of rating and ranking data.

Analysis methods

Preliminaries

The statistical literature on methods for analyzing ratings and rankings seems boundless. It is of very limited interest in this note, to the mere extent that it suggests a small number of methods within reach of humanitarian data analysts. Chiefly, this means that they can be performed with the most popular workhorse application, MS Excel. In addition, we describe and exemplify a method that so far has been implemented only in a statistical program. We introduce it because it delivers a ratio-level priority measure on the basis of needs rankings. Similarly, references to the literature are occasionally useful and necessary for pointers on the validity of some practices that we might otherwise take for granted.

Terminology

We divide the possible methods by the dominant focus of the analysis and by the type of data, ranking vs. rating. The dominant focus is either on the comparison of cases or on the comparison of indicators. Some readers may be habituated to different terminology for "cases" and "indicators". Instead of "cases", "observations" is a more generic term (they are not synonymous. Cases can have several observations, each with its own record in the database and a common case ID, such as when data are recorded from several sources. This note does not deal with multiple-record situtations.). Some (e.g. branches of decision sciences) may speak of "objects" (the objects to evaluate); the psychometric literature generally refers to "subjects". "Indicators" may be terminologically close to the "aspects", "attributes" or "items" that they capture. "Variables" is the most generic. The confusion will go away if we recall that in standard data tables, "cases" or their equivalents occupy a row. "Indicators" and their kin operate by column.

In needs assessment datasets, cases appear as provinces, districts, communities, sites, camps, and similar, and in later stages, households and individuals. Sectors, clusters, problems, or the particular needs or shortfalls defined in relation to them appear as indicator-defining elements. Other types of entities – e.g., geographical units, humanitarian agencies, relief types – may serve as cases or as variables, depending on context and objective. Usually, their function in the data setup is easy to recognize.

Minimizing confusion

At first reading of method descriptions, confusion may arise between "rankings" and "ratings". Combinations can occur:

- Ratings may be combined in a composite measure. Cases subsequently can be ranked on the composite. The first operation is horizontal (within each row), the second vertical (in the column of the composite).
- Obversely, in a first step rankings may be aggregated over subsets of cases (e.g., by geographic area) (vertical operation). In the second, the ranked items (e.g. sectors) can be compared on the aggregate measures (horizontal, e.g. to create regional needs profiles).

The point to take away is that rankings can be built on some function of ratings, and vice versa. This happens in sequence; variables are not ratings and rankings at the same time.

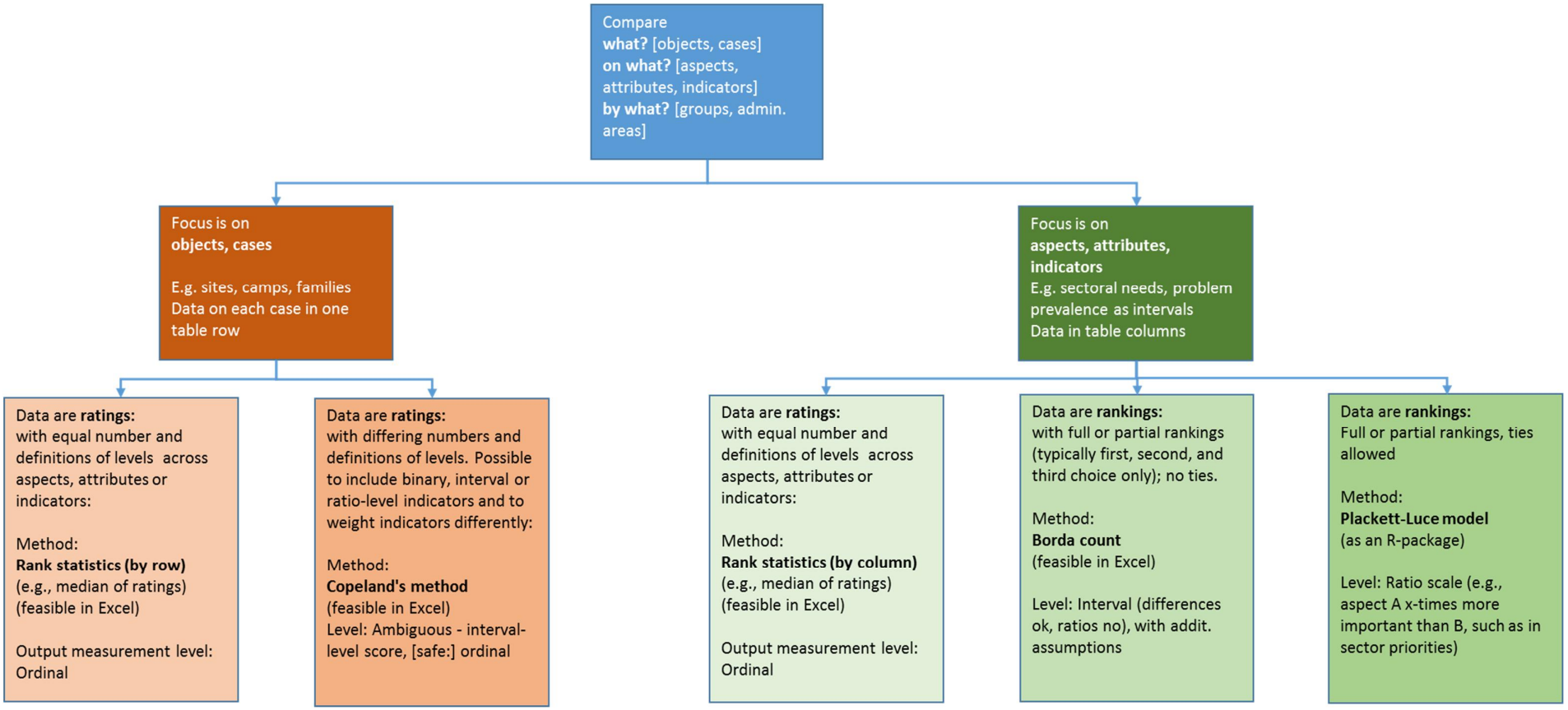
Selected methods

From the panoply of methods, we select four, for the reasons noted above. One of them appears on both the "focus on cases" and the "focus on indicators" sides, making for five analysis situations, each with its particular resulting measures. These are set out in the following schematic.

An additional concern is the capacity of these measures to be broken down by groups (types of affected persons, geographic area, etc.), or not. If they can, are there statistical tests for significant differences, assuming the assessment is based a true probability sample of cases? Such situations may be rare in humanitarian needs assessments; purposive samples and full enumerations are more common. For space reasons, the group-wise and test notes are placed in two charts separately for the "cases" and "indicator" discussions.

We must emphasize that this is not at all a survey of existing analysis methods for rating and ranking-type data. There are many more, and the literature is growing fast. The growth is driven particularly by academic and corporate research into Web-based recommender algorithms. These Big Data-driven efforts are of limited direct interest, but in time they will diffuse some useful new methods or raise established ones to new prominence. For example, the Borda count, surprisingly, seems to be getting more attention in that milieu.

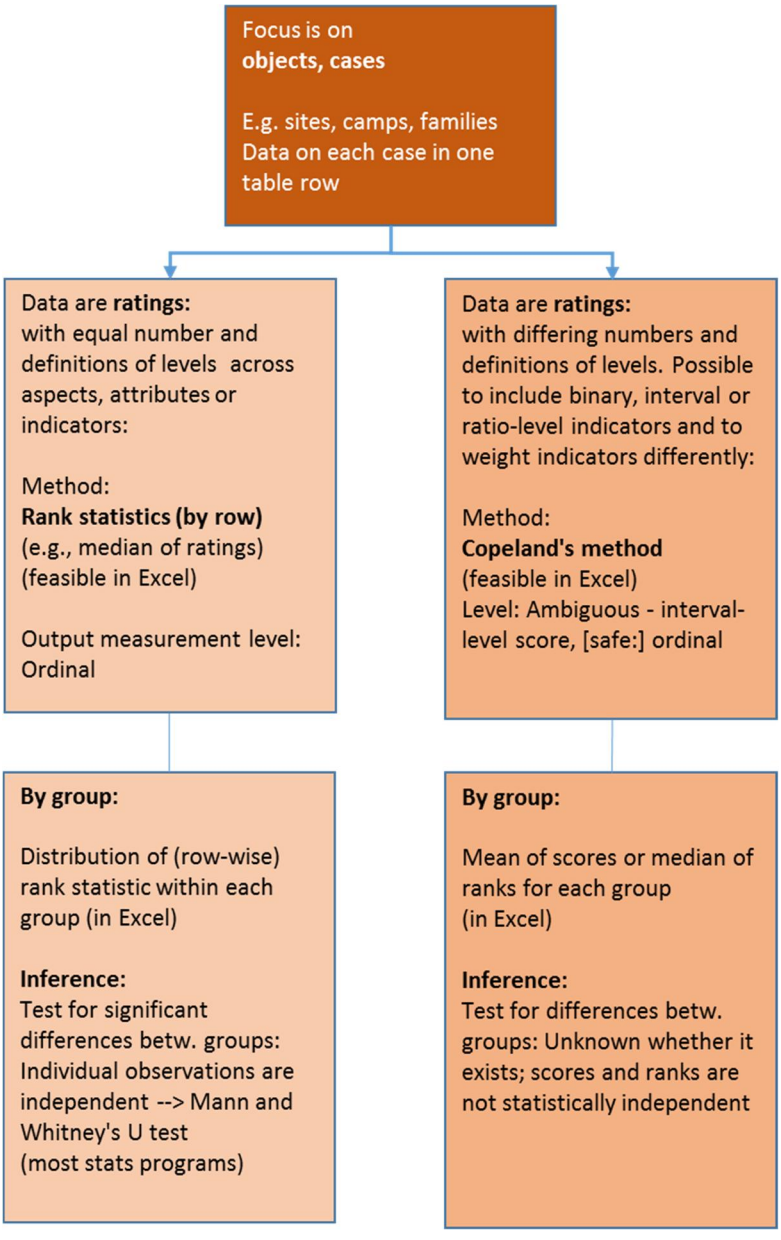
Figure 2: Select methods, by analysis focus and data type



Focus on cases

Methods that aggregate multiple ordinal indicators into scores for individual units work with ratings. They do not work with item rankings (i.e., ranks row-wise, for every case separately). Barring missings and ties, the row sum of ranks is the same for all cases; cases cannot be differentiated on that basis. The situation in which cases are *initially ranked* (column-wise ranks), without first having other indicator information on them (*ratings or metric measures)*, can hardly occur.

Figure 3: Two methods for aggregating ratings



The analysis of rating data faces two different basic situations. In the first, speakers for, or observers of, the cases rate the relevant items on a scale with equal number of levels and identical meanings of the levels (e.g., seven from "totally disagree" to "totally agree" for all items whose ratings will be aggregated). In the second, the number or definitions of levels or both vary across items. The following chart splits over this difference and notes consequences for the choice of aggregation method and of measures of group differences.

Ratings on identical scales

In the first situation, cases are rated across items on identical scales. The ratings are thus intended to be measurements of one and the same underlying construct, such as a one-dimensional deprivation quantity. This one-dimensionality may be ascertained from prior studies or by exploratory analysis of the data

on hand. Or, perhaps more often, it may simply be taken for granted. The construct can be approximated by a rank statistic, i.e. a percentile value of the item ratings recorded for a given case. Depending on assessment interests, this would naturally be the median (e.g., to characterize an average level of deprivation across the concerned sectors) or the maximum (to signal the highest level of deprivation that the unit is experiencing in any sector).

How serious is upward bias?

However, in humanitarian situations, informants have an incentive to upwardly bias descriptions of deprivation, in order to influence local relief allocations (in protection matters, the bias may operate downwardly, underreporting problems for fear of persecution). Overall, the upward bias is likely to push ratings to the upper levels, but the detailed structure of the bias is unknown. Plausibly, some informants remain "objective" in all their judgments. Others exaggerate the degree of unmet needs in some sectors or for some groups of affected persons in their localities towards which they wish to draw privileged attention. Yet others will choose high and extreme ratings "across the board".

Without more knowledge of the biases, it is difficult to recommend any particular rank statistic as the best suited to produce the composite measure. The maximum rating is particular sensitive to upward bias. The median is more robust only if exaggerated ratings are few, i.e. minority outliers. Informants who inflate half or more of their ratings cause inflated medians.

There are some avenues one may pursue in order to gauge the extent of the upward bias in severity ratings. They are purely exploratory. If the median and maximum ratings are on the same level for a low proportion of all cases, one would think that bias is mild. Similarly, the distribution of the median ratings can be seen in the light of common knowledge about conditions in the affected population. An illustration comes from the Rohingya refugee camps in southern Bangladesh.

In March 2018, the International Organization for Migration (IOM) conducted Round 9 of its Site Assessment in all the camps. In the smallest administrative units, in 1,807 "blocks" or "camp points", interviewers elicited severity ratings from block elders on 17 areas of need. The five levels ran from "Not severe" to "Extremely severe". The elders supplied complete, i.e. 1,807 x 17 = 30,719 ratings. This table shows the distributions, over the five levels, of all ratings, block-wise median ratings and highest ratings.

Table 3: Distribution of severity ratings in Rohingya refugee camp units

Level	Meaning	All ratings (*), percent, by level	Blocks, percent, by median rating	Blocks, percent, by highest rating
	N	30,719	1,807	1,807
1	Not severe	3.7	0.0	0.0
2	Somewhat severe	16.1	6.6	0.0
3	Moderately severe	30.9	49.1	0.1
4	Very severe	30.7	42.9	14.4
5	Extremely severe	18.6	1.4	85.5
Total		100.0	100.0	100.0
(*) 1,807 camp blocks * 17 sector ratings per block informant				

By cross-tabulating medians and maxima, we find that the median and maximum ratings are the same for 4.76 percent of the blocks (25 blocks at level 5; 61 blocks at level 4). This proportion is low, arguing against widely inflated ratings. Moreover, in almost half of all blocks, the elders rated the overall situation "somewhat or moderately severe" (56 percent of the block medians); medians at "very or extremely severe" made up 44 percent. Given the precarious living conditions in the camps, as known from media and agency reports, an outsider is tempted to think that the block elders' rating behavior, if anything, understated the true severity.

The ratings median as a measure characterizing cases

We take two points from that illustration:

- It suggests that the median of the severity ratings differentiates cases (here: camp blocks) informatively. The resulting distribution has practical value by virtue of its consistency with other knowledge and of the relatively small groups at both ends of the ratings range ("somewhat severe" and "extremely severe").

- Assuming that the levels sounded ambiguous in the ears of many respondents (what is "somewhat", what is "moderate"?), there must be considerable uncertainty in deciding between adjacent levels. This is a source of measurement error. In the further course of analysis, one may consider amalgamating adjacent levels. Thus the median rating variable could be simplified to a dichotomous measure, with "very or extremely severe" marked as "severe", and "somewhat or moderately severe" as "less severe" ("not severe" median values do not occur in the Bangladesh dataset). Such a reduction loses information; its benefits may outweigh the loss. The refugee population living in blocks with "severe overall conditions" is easy to calculate and cross-tabulate by all sorts of categorical variables of interest.

Comparisons by group

First a step back. We tabulate cases by the medians of ratings, and further break them down by another variable of interest, plus a statistic of interest from yet another variable. In the Bangladesh example, camp blocks are counted by medians of the severity ratings. They are broken down by arrival periods. The additional statistic is the sum of block populations living in the cross-table cells and margins. The question of interest is whether those among the refugees who have been in the camps for longer times have somewhat improved their conditions relative to the newer arrivals.

Table 4: Describing a population by the median of ratings and by additional variables

Arrival period of the majority of the camp point population	Median sectoral severity rating				
	Somewhat severe	Moderately severe	Very severe	Extremely severe	Total
Before 25 Aug 2017	18	106	81	1	206
	9386	56621	50455	487	116949
After 25 Aug 2017	95	687	584	21	1,387
	45624	338783	290937	12631	687975
Unknown for majority	6	94	111	3	214
	2814	43029	46444	1101	93388
Total	119	887	776	25	1,807
	57824	438433	387836	14219	898312

To get a better grip on the data for answering our question, we calculate the row percentages. For simplicity, we give only the population shares, not the block counts.

Table 5: Population shares by camp point severity ratings, by arrival periods

Arrival period of the majority of the camp point population	Median sectoral severity rating				
	Somewhat	Moderatel	Very seve	Extremely	Total
Before 25 Aug 2017	8.74	51.46	39.32	0.49	100.00
After 25 Aug 2017	6.85	49.53	42.11	1.51	100.00
Unknown for majority	2.80	43.93	51.87	1.40	100.00
Total	6.59	49.09	42.94	1.38	100.00

The differences between the two groups with known arrival period are minor. Because the site assessment covered all blocks, the question of test is moot for any result.

In sum, the row-wise median of ratings with identical number and meanings of levels is a plausible measure to characterize cases by the overarching concept (e.g. deprivation) to which the items speak. Yet it must be understood that it is and remains an ordinal measure, and only operations legitimate at this measurement level should be pursued[4]. Of greatest interest is the distribution of the medians by the categories of one or several grouping variables, as demonstrated in the above tables.

[4] Primarily rank statistics, of which the median is the most useful. Metric-based averages, such as the arithmetic and geometric means (the latter enjoys favor among humanitarian IMOs) are not legitimate. For an example of why the geometric mean of severity ratings is nonsense, see Benini (2016:24). Legitimate alternatives that convert ratings to metric scales, such as the ridit, are discussed in the sidebar below.

[Sidebar:] Converting ratings to metric variables

The treatment of ordinal data with methods that require metric data is to be discouraged; it resembles attempts, known from fairy tales, to spin straw into gold. However, just as straw can be woven into useful items such as baskets, the distributional information in ordinal data can be marshalled in ways to form metric variables with some justification. The new metric does not meet a "gold" standard; the ordinal "straw" is still visible. In particular, the transformation depends on the sample of units and items included; the values will be different for other samples.

Signaling for attention

The ordinal data may need several transformations in order to yield a useful construct. By "useful", we do not mean empirically validated, but rather capable of signaling cases that need priority attention. The signals should be more sensitive to higher ratings than the median and more sensitive to lower ones than the maximum. The median rating is a partially compensatory measure; the maximum is non-compensatory5. We look for a fully compensatory one that registers any change in the distribution of the ratings, although, and deliberately, in non-linear fashion.

Also, the construct should be more finely grained than the rank statistic, which has the same theoretical number of distinct values as the rating scale has levels. This table gives an intuitive illustration; the example values of the desired metric are arbitrary, except for the sensitivity to higher ratings, as seen in B < A < C < D, but D < 5.

Table 6: Ratings aggregated in rank statistics vs. a metric measure

Case	Ratings	Median	Maximum	Expl. metric
A	(1,1,1,1,5)	1	5	3.5
B	(3,3,3,3,3)	3	3	3.0
C	(3,3,3,3,5)	3	5	4.0
D	(3,3,3,5,5)	3	5	4.3

The comparison between A and B is instructive. On the median, B > A. On the metric construct, A > B. Under the rules for ordinal statistics, arbitrary monotonic transformations of the data are allowable. But conversions from ordinal to metric that provoke rank reversals in aggregates are not encouraged; Cliff, in *"What is and isn't measurement"* does not want us to be "at the mercy of whimsical but 'legitimate' transformations" (1993:72).

However, *selective* reversals of the median-based order may be the very outcome that the metric conversion is expected to produce. These deviations from the order that the median ratings produce are desirable because of their signaling value. They bring to attention cases that have one or several high ratings, some of which may be below the scale maximum.

Order of transformations

The production of the metric measure may go through several transformations:

[5] Compare these sets of four ratings each on their medians and maxima: (4,4,4,4), (3,4,4,5), (3,3,5,5) and (2,4,5,5), all with the same ratings sum.

1. **Data-driven:** The distribution of ratings over all cases and items of interest determines a first raw interval-level transformation. Below we demonstrate the ridit-transformation, which has several useful properties.
2. **Theory-driven** (optional): The second transformation creates new values that emphasizes or de-emphasize the effect of the higher ratings on the ultimate composite measure. Some types of transformations may have properties that facilitate subsequent analyses. At this stage, the analysts should determine whether the new construct should be considered ratio-level (it has a natural zero point such as when the lowest rating may be taken to mean that there are no unmet needs in the particular needs area - e.g., NFI distributions covered all households with all essential items).
3. **Policy-driven:** Policy considerations may determine parameter settings of the transformation under #2. For example, if key informants use the highest level of the rating scale to signal life-threatening deprivations, the transformed values for that level should be such that one item with a top rating is enough to cause a high value of the aggregate. By contrast, if the highest rating expresses greater scarcity in a more settled situation, then the escalation in the transformed values should be less steep.

At this point, we calculate the composite measure, as a row-wise statistic of cases, and/or a column-wise one for items. Since this is an (at least) interval-level conversion, and the non-linear transformation has already taken place, the arithmetic mean is the statistic of choice.

4. **Rescaling** (optional): For easy comparison, the composites can be divided by their theoretical maximum and, if desired, multiplied by the number denoting the highest level of the original ratings. Thus, for a five-level rating scale, the metric scale can have its theoretical maximum rescaled to 5. In addition, rescaling the minimum to 1 implies an interval-level interpretation; if ratios are defensible, the minimum should not be rescaled.
5. **Population-weighted** (optional): The usual reservations apply, particularly when ratings are elicited from only one key informant per unit.
6. Weighted by **item trade-offs** (optional): The transformations to the (at least) interval level values and their aggregation by the arithmetic mean imply that true importance weights are not available. The use of trade-off weights may please analysts, but it second-guesses the original raters.

Before we propose specific transformations, we must spell out two tacit assumptions that the conversion from ordinal to metric makes:

- **Ratings are comparable:** The assumption is that the meanings of the rating levels are the same for all raters and for all rated items. This is as unlikely as it is indispensable. The differences in individual understandings may nearly cancel out across raters for a particular item, but they will likely defy comparability across humanitarian needs areas. The meanings may be sufficiently similar for needs areas that call for physical relief (food, shelter, etc.), plausibly including health care. Physical and protection needs may be harder to accommodate on a scale with levels that can be meaningfully equated.
- **The categorical ratings map to a continuous variable:** The metric conversion implies a continuous underlying construct. To understand why, it is sufficient to

note that for a rating scale of *k* levels applies to *j* items, rank statistics such as median and maximum can take only a small number of distinct values (no more than *k* for the maximum and, for odd numbers of ratings, the median; 2*k* – 1 for the median for even numbers). The arithmetic mean of the converted scale can take many more. The metric scale is much more finely grained. Intuitively, a continuous construct is an appealing idea for the measurement of physical deprivation. For example, other things being equal, the extent of food deficits should be monotonically related to excess mortality. The idea of continuity is much less helpful when we deal with levels of threat and violence. Transitions such as from harassment to expulsion, from detention to killings, etc. constitute categorical breaks. It does not seem appropriate to map them to a continuous variable. In such situations, one may have to be content with the ordinal scale[6].

These assumptions are problematic. If we wish to go ahead with the metric conversion, we cannot avoid them.

The ridit transformation

The ridit is a data-based transformation (Benini 2016, Wikipedia 2016b). For a variable with k = 1,2, .., *i, .., k* ordered categories, its value for category i is defined as ½ of i's sample frequency + the cumulative frequency up to the category next below, $i - 1$. As long as the frequency of the lowest category is > 0, the ridits, by construction, fall into the interval $0 < r_i < 1$. This has several advantages. As a frequency, the ridit formally is a ratio-level variable (because it is an expression of several probabilities). At the same time, the value for the lowest category is not tied to zero, but is half of frequency of the lowest category; thus no need to justify that the lowest category has always zero effect on the aggregates. At the upper end, the ridit decreases with the frequency of the last category; this is particularly helpful with upwardly biased ratings. The more the top category is overused, the less its "weight" on the metric scale.

This small example demonstrates the ridit transformation for the five-level severity scale. Suppose that all the ratings in the data table whose distribution is to define the ridit are in a range named "ratings". Using the data in the ridit demo worksheet in the companion Excel file, we arrive at this table:

Table 7: Calculation of the ridit

	1	2	3	4	5	6
1	**Scale, verbal:**	**Not severe**	**Somewhat severe**	**Moderately severe**	**Very severe**	**Extremely severe**
2	Scale, numeric:	1	2	3	4	5
3	Frequency, all ratings	0.021	0.071	0.407	0.343	0.157
4	Cumul. frequency	0.021	0.093	0.500	0.843	1.000
5	Ridit	0.011	0.057	0.296	0.671	0.921

with these formulas in rows 3 – 5:

Frequency, all ratings	=COUNTIF(ratings, R2C) / COUNT(ratings)

[6] Other research areas, such as life satisfaction studies, struggle with the relationship between continuous "dimensions" and abrupt qualitative differences (De Boeck, Wilson et al. 2005, Schröder and Yitzhaki 2017).

Cumul. frequency	=SUM(R3C2:R3C)
Ridit	=R3C / 2 in R5C2, and =R3C / 2 + R4C[-1] in R5C3:R5C6

The number of possible further transformation functions is infinite, but only a few are of practical interest. We rescale their maxima to 1 so that they can be compared to the cumulative frequency and to the ridit:

Table 8: Further transformations of the ridit

	1	2	3	4	5	6
7	**Ordinal scale and ridits:**					
8	Scale, numeric:	1	2	3	4	5
9	Cumul. frequency (copy)	0.021	0.093	0.500	0.843	1.000
10	Ridit (copy)	0.011	0.057	0.296	0.671	0.921
11	**Possible transformations of the Ridit (selection):**					
12	**Raw:**					
13	Logit	0.011	0.061	0.421	2.043	11.727
14	*Formula:*	=R10C / (1 - R10C)				
15	Power function, base = 10	1.025	1.141	1.979	4.693	8.345
16	*Formula:*	= 10^R10C				
17	Power function, base = 15	1.029	1.167	2.232	6.161	12.125
18	*Formula:*	= 15^R10C				
19	Exponentiation	1.011	1.059	1.345	1.957	2.513
20	*Formula:*	= EXP(R10C)				
21	Softmax function	0.128	0.134	0.171	0.248	0.319
22	*Formula:*	= EXP(R10C) / SUM(R19C2:R19C6) [sums to 1]				
23	**Rescaled so that maximum = 1:**					
24	Logit	0.001	0.005	0.036	0.174	1.000
25	Power function, base = 10	0.123	0.137	0.237	0.562	1.000
26	Power function, base = 15	0.085	0.096	0.184	0.508	1.000
27	Exponentiation	0.402	0.421	0.535	0.779	1.000
28	Softmax function	0.402	0.421	0.535	0.779	1.000

The softmax function is mentioned here only because it nicely sums to 1 (Wikipedia 2019).

The sequence of ridits, from a low base of 0.01 to the top value 0.92, represents a relatively modest escalation. The ratio between “very severe” and “extremely severe” is less than 1 : 1.4. Thus, when choosing a transformation of the ridit – if the ridit itself does not serve the purpose -, we should commonly ask two questions: Should the scale start from a low base (with these data: logit and the two power functions), or from a higher one (exponentiation)? And: how strong should the transition be from the second highest to the highest category, expressed as ratio between the two transformed values? (with these data: almost 1 : 6 for the logit, roughly 1 : 2 for the two power functions, less than 1 : 1.5 for exponentiation).

The logit may be appropriate when the real-life situations that prompt key informants to opt for “very severe” vs. “extremely severe” are dramatically different, and the ratings are fairly reliable, with minor upward bias only. Absence of bias minimizes false positives in the diagnosis of extremely strongly affected units. If either of those conditions does not

hold, a power function may serve better. If, for whatever reason, one must assume that the underlying construct should be fairly elevated even for the lowest categories, exponentiating the ridits might do.

On the brink of alchemy

It should be clear by now that the conversion of ordinal to metric scales is problematic. First of all, no new information is gained. All the arithmetic mean as the aggregation function on the transformed values does is to create more distinct values, thus making the composite measure more finely grained[7]. This is helpful. Second, the ridit follows deterministically from the frequency distribution of the ratings. However, using it or any further transformations as the basis of the metric scale relies on value judgments, analytic interests and beliefs about the data (e.g., the extent of upward bias). Third, one such value judgement is implied in the acceptance of rank reversals, that is, accepting that for two units A and B, A has a higher median of the ratings, but B is higher on the mean of the transformed scale.

Thus the validity of the resulting scale or scales is problematic and hard to establish empirically. The pragmatic question is whether we can do without these constructs. Rank statistics of severity scales are rather dull – coarsely grained and insensitive to most smaller changes. The scope of statistical operations permitted, and the number of procedures suited, for ordinal variables, is limited. Conversion to metric levels widens the analytic horizon and, specifically for Excel users, makes more of the commonly used functions legitimate. Like the alchemy of earlier ages, modern scale transformations call for inventiveness and bravery, and then for sober assessment of what blends together in the heated crucible.

Ratings on different scales

Situations may occur where cases are rated on several items for which the rating tool provides different scales. The differences may be in the number of levels and/or the meanings of levels.

Illustrative scenarios

- **Different numbers of levels:** By way of example, assume that a health sector coordinator wishes to evaluate coverage of a number of refugee camps. For the various services that the sector agencies are providing a simple three level scale is developed: “not available / sometimes available / always available”. Given enough items, three levels may be adequate in order to roughly gauge the availability of services that clients can access during normal business hours. It would not be enough for emergency services, say, ambulance transport, for which the third level has to be refined into something like “available during working day business hours only” and “available 24/7”.
- **Same number, but different meanings of levels:** Going back to the Bangladesh site assessment, the same five-level scale of severity was applied to needs for goods and services as different as food, water, shelter, education, psychosocial support,

[7] The same would be true if we estimated an underlying metric construct with advanced statistical models, such as structural equation modeling.

job opportunities and several more. If a respondent used the level "extremely severe" for the need for food as well as for the need for education, its meaning can hardly have been the same. Extremely severe lack of food threatens health and survival in the short term; the breakdown of education has severe consequences in the longer term. While unmet needs in all those domains and service areas ultimately belong in a broadly defined deprivation concept, combining ratings for all of them in one measure threatens its validity. What does this quantity effectively measure?

- **Different numbers as well as different meanings of levels:** Imagine a country hosting a large refugee population in the perspective of prolonged stay and active social integration in terms of housing, education and employment. Accessibility of public transport is an important factor of social inclusion, for refugees as well as for the poorer segments of citizens, and indirectly for the harmonious relations between the two groups. Suppose a study of social inclusion takes place. It interviews samples households in mixed and segregated neighborhoods, with a special module on transport accessibility. The researchers create an instrument that measures both perceived access and some of its objective correlates:

    - For the former, they modify Lättman's four-item "Perceived Accessibility Scale"[8] (Lättman, Friman et al. 2016), maintaining its form to that point.
    - They augment it with items on bus service reliability, comfort and behavior towards customers. Sample households with members traveling on buses are asked about typical on-board travel time (in minutes, coded to five intervals), walking time to the bus stop from their homes (same form), number of transits between home and workday destination (counts, observed range: 0 – 3), estimated average minutes late departure (five intervals, but different from above; zero minutes is a valid category).
    - Researchers also check the bus stops most used by the respondents for the presence of printed schedules. Some bus lines have them in the local language, others have larger versions with one half in the refugees'. These observations are coded "None or damaged / local language only / bilingual". Also, it is noted whether any bus companies serve the neighborhood on weekends – coded as simple "yes" and "no".

Simple adjustments

It is obvious that in none of the three situations can the row median of the various indicators be used as the intended measure with at least some face validity. In the first – services within the health sector -, some mild trickery could be forgiven "to make things work". The three-level items could be recoded by raising "always available" to the same level as "available 7/24" in the four-level ones, in other words from "3" to "4". This seems defensible if clearly stated. In the second situation, items about needs gaps with delayed

[8] The original items are: 1. "It's easy to do [daily] activities with public transport", 2. "If public transport was my only mode of travel, I'd be able to continue living the way I want", 3. "It's possible to do the activities I prefer with public transport" and 4. "Access to my preferred activities is satisfying with public transport", rated on seven levels from "disagree" to "completely agree" (ibd., 39). Item no. 2 is the least suitable for talking with refugees.

consequences should be removed from the composite measure of short-term severity. In such situations, aggregation may be formally straightforward, but the validity of the item mix is at issue.

Difficult indicator constellations

The third situation is the truly interesting one – the one that justifies adding a new method to the toolbox. Not only do these items come with different levels, from two for weekend service to seven for subjective perception, but any reduction or expansion to the same number of levels for all the items would not ensure comparability. Moreover, some of the items are metric (walking, waiting, riding in minutes, transits between lines); recording them in categorical ranges loses information and analytical flexibility.

What is needed in this and analogous indicator constellations is a method that incorporates items with any number of (strictly ordered) categories as well as binary, count and continuous indicators. One of the methods that achieves this is known as "Copeland's method" (Saisana 2016, Wikipedia 2016a), sometimes called the "Copeland Rule".

Copeland's method

Capabilities

Copeland's Method, like the Borda Count discussed below, is essentially an election method (Wikipedia 2016a). As such, it produces a ranking with one or several winners. However, it has been studied and applied in social choice and multi-attribute decision making. Saisana and Saltelli reviewed it, alongside other methods, in the context of ratings and rankings (Saisana and Saltelli 2011). Copeland has several strengths that make it attractive in the ratings context. In particular, it takes care of the above-described third situation – combining indicators of different measurement levels, including ratings, and with different numbers and meanings of levels. Saisana (2016, op.cit.) enumerates several other benefits. Copeland has

- No need for outlier treatment
- No need for data normalization
- No need to attach monetary values to indicators (which says, in the ratings logic, no need for common meanings of the levels)
- No need for data aggregation (additive or multiplicative or other functions), and
- Allows for true importance weights (as different from mere trade-offs in compensatory methods)

In other words, Copeland's is a non-compensatory method. A decrease in one indicator cannot be compensated for linearly by an increase in another, at a rate fixed by their relative scales and weights[9].

[9] For a quick and accessible, if outdated, overview of non-compensatory methods, see Yoon and Hwang (1995); for a more up-to-date and technical discussion, see Greco et al. (2016).

The downside of non-compensatory

However, this opens the method to criticisms. Copeland is based solely on the sign of the differences on given criteria, i.e. Unit A has a higher, identical, or lower value on criterion *X* than Unit B. The size of the difference $x(A) - x(B)$ is not taken into account. This property will at times contradict humanitarian rationality. Suppose we compare two camp populations, A and B, on three WASH criteria – the estimated proportion of households with enough water storage vessels, whether the latest distribution of hygiene articles took place more than two weeks back, and the amount of drinking water available per day and person. Assume that A is in a better position than B by the first two criteria. The water situation in camp A is desperate, with only one liter per person per day tankered in. The people in B receive considerably more. In this narrow perspective, the Copeland Rule says that B's needs are greater than A's. From a life-saving values orientation, that is not right.

The mechanics

Yet, one should not dismiss Copeland from humanitarian analysis too readily, given its ability to deal with mixed data situations. To achieve this, Copeland breaks with the logic of constructing case-based measures that only use the individual case's own values in the variables to be aggregated. Instead, the relevant indicator data on *all cases* are used in generating the score of each case. This happens in three steps:

- First, the indicator values of every case are compared to those of every other case. For each pair of cases A and B, A receives an outranking score w(A vs. B) = the sum of the normalized weights of the indicators on which A dominates B.
- Second, the outranking scores are transformed into marks of +1, 0, or -1. For A vs. B, the mark for A is +1 if the sum of importance weights on the indicators on which A dominates B is larger than for those on which B dominates. If A's weight sum is smaller than B's, A receives -1. The mark is 0 if A and B are tied.
- Third, the final Copeland score of a case is the sum of its marks that it received from the comparisons with all other cases. Normally, the cases are then ranked by these scores.

Indicators with the same orientation

For this method to work consistently, the indicators must have ordered values. Their orders must all have the same substantive direction or, as some might say, orientation. That is, for all, either "greater means worse" or "greater means better" must hold. To this extent, the indicators are indeed under a common overarching concept (e.g., higher values imply more severe deprivation) even though the number and meanings of their levels may differ. Some indicators may need reverse-coding in order to conform to the orientation of the rest.

Scores and ranks, tests

Two more remarks seem in order:

- **Using the scores, not only the ranks:** Because Copeland's Method was developed in the context of ranking candidates in an election, there seems to be an interest in the scores only as the precursor of the ranks, not for their own information value.

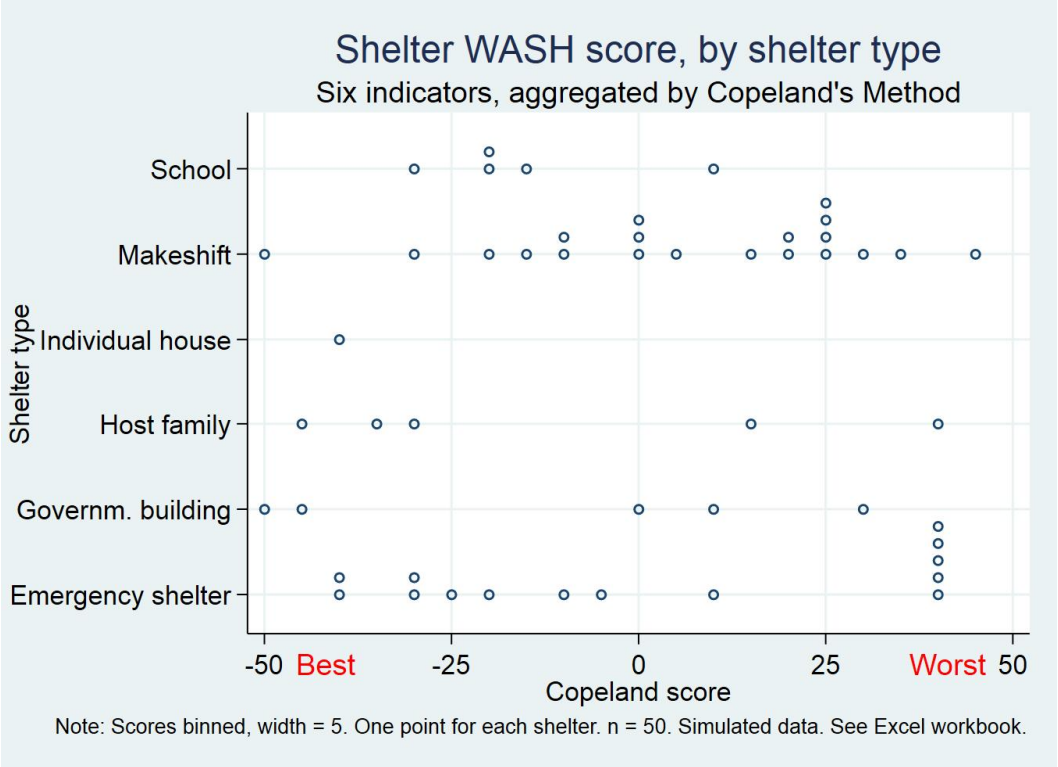
The Copeland score, being the difference between two counts (wins minus defeats in the pairwise contests), is an interval-level variable. On such a variable, the arithmetic mean is a legitimate operation (analogously to "average temperature").

- **Tests for differences among groups:** Whether such differences among groups are statistically significant (assuming that the units are a random sample), and which tests would be valid to establish significance, are difficult questions. The difficulty arises from the fact that the observations (the indicator values) are independent, but the Copeland scores are not. The score of a unit incorporates information about this unit as well as all others, as we have seen from the way it is calculated. In these circumstances, it is unknown whether classical tests for group differences– Wilcoxon's rank-sum, Mann and Whitney's U, and the non-parametric equality-of-medians tests – are applicable, and if so with what corrections for statistical dependence.

Exploration before testing – if any

The problem may be academic. Most samples in humanitarian assessments are purposive. A careful exploratory analysis may be more insightful than questionable tests of differences among groups of one parameter (median or mean). In the dataset simulated for the demo workbook, the WASH conditions in IDP shelters is worse for those in makeshift shelters than for all other types combined. The median Copeland scores confirm that. However, a multiple histogram of the 50 cases quickly gives away that the variance is more important than the central tendency (see graph). In particular, among 14 emergency shelters, a subgroup of five stands out in distinctly worse conditions. In real life, managers would want to know what factors account for that, and whether this justifies extending the sample in this particular and/or other types. Some shelter types are represented so scarcely that even guesses about the distribution in the wider population are impossible. An extreme score may reflect multiple errors in the data on the same case. The scope for formal tests of group differences is very limited. Deep inspection matters.

Figure 4: Variability of WASH conditions, by IDP shelter type



Implemented in MS Excel

Saisana was the first to develop an adaptable Excel implementation of the Copeland Method. It is part of the *"COIN tool: A do-it-yourself guide in Excel for constructing and assessing composite indicators"*, a resource offered by the European Commission Joint Research Center in Ispra, Italy (Saisana 2015, Saisana 2016)[10]. Our revised version simplifies the worksheet architecture and particularly the formulas employed in the so-called outranking matrix. With both versions, though, the user needs to be comfortable with some mid-level Excel tools. Our version relies on named ranges and the versatile INDEX function, which make formulas shorter and more transparent.

Focus on indicators

Objectives

When we wish to compare a collectivity of cases on several comparable attributes, the focus of the investigation moves to the indicators. Objectives vary; we may want to compare the distributions of attributes that the indicators express, or find significant associations among them, or test for associations with external variables of interest.

[10] Training presentations are available at https://composite-indicators.jrc.ec.europa.eu/?q=past-events , but there does not seem to be a link to the COIN tool Excel workbook itself.

- The first of those objectives is the most common in humanitarian analysis, notably to establish priorities among sectoral needs. Similarly, "problems" in the entire population can be rated on the bases of prevalence estimated in local units.
- The second and third objective both necessitate forms of correlational analyses. The second lets us look into how levels of unmet needs or unresolved problems go hand in hand across different sectors.
- In the third, levels of needs or problems are cross-tabulated with categorical attributes or regressed on categorical and/or continuous ones.

Commonly in humanitarian analysis, the outside attributes are demographic, spatial, temporal or aspects of the humanitarian action. The more we subdivide the population into groups, the more the focus on indicators shades into the focus on cases.

Robustness

In most needs assessments, the number of comparable indicators is smaller than the number of cases. At first sight, that appears to make indicator comparisons more robust than case comparisons. Suppose the assessment sample includes 100 affected communities and, besides water and food, ten more needs areas. First, we calculate the strength of association between levels of unmet needs for water and for food measured on rating scales. Second, we compare two particular communities X and Y on all twelve needs areas.

Instinctively, we will give the first result, based on 2 * 100 = 200 data points, more credence than the second, which considers a mere 24. This confidence is not necessarily warranted. Analyses that focus on indicators are as vulnerable to problems of reliability and validity as those focused on the cases. In the extreme, it could happen that the data on X and Y were highly reliable while the other 98 communities had bad data. The comparison of X and Y would still be instructive. The correlation of the water and food needs ratings would be worthless.

Relevance

Indicator-focused analyses are justified, not by their presumed robustness, but because of their relevance to humanitarian priorities. The humanitarian response is structured primarily by the mandates, competencies and resources of sectoral agencies. Needs are interdependent, but their measures are more closely associated with some sectors than with others. Multi-sectoral assessments seek a combined view, within the social and spatial pattern of affected communities. The indicator-focused analysis helps to clarify sectoral priorities. This should take place *before* sector agencies translate the information by the specific technical codes that make the response practical.

We assume that the indicators needed for prioritization arrive as ordinal variables, as ratings or rankings (prevalences estimated as ranges may be treated as ratings). We discuss three situations, as shown in this diagram.

Figure 5: Three methods of indicator-focused analysis

Data are ratings

Ratings are statistically independent observations. This permits the use of rank statistics both row-wise (for a measure of cases) and column-wise (for statistics of the indicators). This is convenient because the Excel functions MEDIAN and MAXIMUM work in both directions. For indicator statistics by group, those can be combined with the IF-function (the standard Excel Pivot table does not compute medians).

For illustration, we return to the NPM Round 9 data on 1,807 Rohingya refugee camp points mentioned earlier. This table presents the distribution of severity ratings in 17 needs areas. Because of large differences in the camp point populations, the ratings were population-weighted.

Table 9: Severity ratings in 17 needs areas, Rohingya camp points, March 2018

	1	2	3	4	5	6	7	8
1	Need	Severity rating					Total	Median rating
2		1	2	3	4	5		
3	Cash	0.2	1.8	5.0	30.0	63.0	100.0	5
4	Firewood	1.2	0.7	2.8	37.5	57.9	100.0	5
5	Water	1.8	5.1	15.5	34.1	43.4	100.0	4
6	Jobs	0.6	5.1	14.8	43.5	36.2	100.0	4
7	Shelter	1.3	5.7	23.6	42.7	26.7	100.0	4
8	Food	0.8	5.7	24.0	48.4	21.1	100.0	4
9	Security	1.7	10.5	28.8	41.3	17.7	100.0	4
10	Education	3.5	9.8	34.4	37.0	15.3	100.0	4
11	WASHs	2.4	15.5	40.6	31.6	9.9	100.0	3
12	Health	1.9	10.6	42.3	35.9	9.2	100.0	3
13	PSS	6.1	30.4	34.8	20.9	7.9	100.0	3
14	Hygiene	4.2	30.7	46.0	13.0	6.1	100.0	3
15	Transport	7.1	31.0	43.0	15.1	3.8	100.0	3
16	Training	5.6	32.5	44.4	14.2	3.2	100.0	3
17	Utensils	12.8	22.5	36.3	25.2	3.2	100.0	3
18	Registration	9.2	27.4	36.8	24.8	2.0	100.0	3
19	Clothing	6.6	30.3	44.4	18.0	0.7	100.0	3
20	*All ratings*	*3.9*	*16.2*	*30.4*	*30.2*	*19.3*	*100.0*	
21	Note: Population-weighted. Row-wise percentages. Sorted descendingly on percentages of rating 5.							

The median ratings are easy to identify by just looking at the row-wise ratings frequencies: but if one is willing to bother with some code and formulas, they can also be calculated: 5 (extremely severe) for cash and firewood, 4 (very severe) from water down to education, and 3 (moderately severe) for the rest of the needs areas[11]. The values of the maxima are not informative; they are the same for all needs areas (5), but the proportions of refugees in camp points with the highest rating in a given sector are of interest.

[Sidebar:] Calculating population-weighted ratings and their medians

Excel users obtain the frequencies using the functions SUM and SUMPRODUCT. If the names of the column ranges in the ratings data table are identical to the needs areas of the table above ("Cash", "Firewood", etc.), the function INDIRECT makes the calculations even easier. Assume the population column range is named "Population", then the row frequencies are produced by

= SUMPRODUCT(--(INDIRECT(RC1)=R2C),Population) / SUM(Population)

Note the double dash (minus sign) before (INDIRECT ..). Note also the mixed-type cell references.

---

[11] The median is the rating at which the cumulative row frequency first turns > 50 percent (if it is exactly = 50 percent, add half of the distance to the next higher value, i.e., for adjacent rating levels, add 0.5). E.g,, for water, the cumulative frequencies are 1.8, 6.9, 22.5, 56.8, and 100 percent. The median thus is 4.

Most users will be quick to obtain the median ratings in column 8 manually, by simple visual inspection of the ratings frequencies. If one is willing to bother with some code, the medians can be obtained as follows:

1. In the workbook with the above table, press Alt + F11 to accesss the VBA Project. Insert a module, or work with an existing one. Insert the following code:

```
Public Function generateCumulativeArray(dataInput As Variant, _
                                        Optional fromValue As Long = 0, _
                                        Optional toValue As Long = 0) As Variant

Dim i                   As Long
Dim dataReturn          As Variant
ReDim dataReturn(0)
dataReturn(0) = dataInput(fromValue)

For i = 1 To toValue - fromValue
       ReDim Preserve dataReturn(i)
       dataReturn(i) = dataReturn(i - 1) + dataInput(fromValue + i)
Next i
generateCumulativeArray = dataReturn

End Function
```

The function creates an array with the cumulative frequencies[12]. These are needed to obtain the median rating from the row-wise percentages (the rating at which the cumulative percentage exceeds 50). The array is then passed to the formula in the cell of the same row in column 8 (next point). (If one does not want to get into VBA code, he may calculate the cumulative frequencies in an auxiliary table to the right of column 8).

2. For easier reading of this formula, name the range that holds the rating scale values “scale”. In this table, “scale” is R2C2:R2C6, with 1, 2, .., 5.

3. In column R3C8, type

=IFERROR(INDEX(scale,1,MATCH(50,
generateCumulativeArray(RC2:RC6,1,5),1)) + 1, INDEX(scale,1,1))

and enter it as single-cell array formula, by pressing Shift+Ctrl+Enter. It will appear in the formula bar with {} around it. Copy-paste the formula to the remaining item rows in C8. The resulting medians will not be correct in the event that a cumulative percentage is exactly 50; the formula does not provide for the correction by +0.5 in such cases. For population weighted ratings, this should happen very rarely.

4. Save the workbook in a macro-enabled format, with the extension .xlsm.

This exercise has also a didactic value; it demonstrates the combination of the functions MATCH and INDEX as a lookup tool as well as the syntax of dynamic arrays to calculate some complex intermediate result, which then can be called by other functions. Both situations occur frequently in Excel-based data analysis.

---

[12] Credit goes to user “Vityata”, https://stackoverflow.com/questions/46343797/arrays-cumulative-sum .

A closer look at sectors with the same median ratings makes it clear that comparisons on mere medians are coarse. This measure places the needs for cash and firewood in a category of higher urgency than the others, and the needs in the next seven area higher than the rest. It does not differentiate between needs areas with the same median ratings. Water and education have the same median ratings (4), but their distributions make it safe to venture that the need for water is the more pressing of the two. A sectoral priority measure should be able to differentiate between the two.

Tests for differences

If the assessment covers a random sample of units, tests help us decide which of two needs areas likely has higher priority, given the ratings distributions. Two approaches are plausible. If we consider the rating scale purely ordinal, without any assumptions about the causal forces that push key informants to move their ratings higher, a sign test will work **(Test A)**. If we believe that the difference between "extremely" and "very serious" is much greater in terms of distress than that between "very" and "moderately serious", we discount anything below "extremely serious" (This belief implies the absence of massive upward bias among raters.). We only test for differences in the proportions of this level **(Test B)**.

We draw a random sample of 100 camp points. We derive two binary indicators for water need rating = 5 (extremely severe) vs. all other levels, and analogously for education. The ratings differ as follows:

Table 10: (several tables:) Tests for differences between two rated items

Variable	Not sever	Somewhat	Moderatel	Very seve	Extremely	Total
Water	5	3	16	29	47	100
Education	3	9	28	45	15	100

The sample medians both are = 4 (very severe).

Test A estimates the probability that pairs of water and education ratings randomly drawn from a population with equal medians would have the same rate of excess of water > education over education > water as the pairs in this sample.

```
Sign test

        sign |    observed    expected
-------------+------------------------
    positive |          51        34.5
    negative |          18        34.5
        zero |          31          31
-------------+------------------------
         all |         100         100

One-sided tests:
  Ho: median of Water - Education = 0 vs.
  Ha: median of Water - Education > 0
      Pr(#positive >= 51) =
         Binomial(n = 69, x >= 51, p = 0.5) <  0.0001
```

The same result will be produced by the Excel formula

```
= 1 - Binom.dist(51 - 1, 69, 0.5, TRUE)
= 0.0000438 < 0.0001
```

The probability of drawing such a sample in a population with equal median ratings for water and education needs is less than 1 in 10,000. With high probability, the need for water is more pressing. Note that the test takes into account all the information, not only the proportion of extreme values (level 5). Population-weighting is not an option for this test.

Test B looks only at those proportions, which are straightforward for n = 100: 47 percent "extremely severe" for water 15 percent for education. The observations are unweighted.

```
. prtest WaterExtreme == EducExtreme

Two-sample test of proportions          WaterExtreme: Number of obs =      100
                                         EducExtreme: Number of obs =      100
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
WaterExtreme |        .47   .0499099                      .3721784    .5678216
 EducExtreme |        .15   .0357071                      .0800153    .2199847
-------------+----------------------------------------------------------------
        diff |        .32   .0613677                      .1997214    .4402786
             |  under Ho:   .0654064     4.89   0.000
------------------------------------------------------------------------------
        diff = prop(WaterExtreme) - prop(EducExtreme)             z =   4.8925
    Ho: diff = 0
    Ha: diff > 0, Pr(Z > z) < 0.0001, one-sided.
```

Note that the 95-percent confidence interval around a 50 percent proportion in a sample of this size (100) is [39.8, 60.2] percent, a range of roughly 20 percent. In other words, "the minimum detectable difference" (MDD) for attributes with middling frequencies is around 20 percent, given that sample size. This may be remembered as a rule of thumb for assuming significant differences based on the proportions in one level such as "extremely serious". For attributes with population frequencies closer to 0 or 100 percent, the interval gets progressively narrower; in our example of "education – extremely severe" (15 percent), it is [8.6, 23.5] percent.

Both tests find that the difference is statistically highly significant. Both tests are feasible in Excel. For the sign test, use the function for the binomial distribution. For proportions, guidance on the normally distributed test statistic *z* is easily found on the Web[13]. With a few more devices – the functions IF, COUNT and AVERAGEIF, as well as single-cell array formulas -, it would be feasible to create a table of test results for differences in all pairs of indicators.

But: Not only would this be tedious; most significant differences would be foregone conclusions. Tests should be applied selectively, for reasonably random samples and for comparisons that are substantively relevant, yet hard to make looking solely at descriptive statistics. Meanwhile, the inspection of distributions, facilitated with conditional formatting, should provide a good first understanding of most of the differences that matter.

[13] E.g., at https://www.wikihow.com/Compare-Two-Proportions .

Items attached to subsectors of a given sector

Rating questions can be used in relation to specific problems with a particular sector of interest. The assessment designers may want to sample several well-known or plausibly existing problems at the sub-sector levels. The selection may be the result of importing an assessment tool tested in other contexts, expert beliefs, conversations with affected communities or any mixture thereof. Nevertheless, the mapping of selected items (the problems) to sub-sectors, to some extent, will always be an administrative construct, only partially congruent with local concepts.

The question of interest is whether the ratings of sub-sector problems can be aggregated in a way that says something valid of the relative importance, not of single problems, but of the sub-sectors.

The Water, Sanitation and Hygiene (WASH) sector offers itself, already by its tripartite name, for an illustration of the conceptual and analytic problems. Suppose key informants are asked to broadly estimate the proportion of persons in their communities (e.g., blocks of a refugee camp) for which the statements in this table hold. The enumerators make sure that the informants understand the common scale as "Approx. 0 – 20 %, 20 – 40%, etc.". Items # 1 - 3, 6 – 9 are formulated as achievements, 4 and 5 as deficits. At data entry, the former are recoded; all items have the same direction, with 1 for "an estimated 0 – 20 percent have access / have enough / live in clean areas", etc., up to 5 for "80 – 100%.

In some agency contexts, such questions are known as "screening questions" regarding the situation within the sector of interest.

Table 11: WASH sub-sector items for rating

Item	Sub-sector	Screening questions
1	Water supply	How many people have enough water to drink?
2	Water supply	How many people have enough containers to fetch/store water?
3	Sanitation	How many people have access to a functioning sanitation facility (latrine/toilet)?
4	*Sanitation*	*How many people share sanitation facilities (latrines/toilets) with other households?*
5	*Sanitation*	*How many people live in areas where dumped garbage is frequently visible?*
6	Hygiene	How many people have enough soap or soap substitutes?
7	Hygiene	How many people have access to functioning hand-washing facilities in their household?
8	Hygiene	How many people have access to functioning bathing/shower facilities?
9	Hygiene	How many females have enough female hygiene products?

In this example, the subsectors are represented by unequal numbers of items. That could be corrected by selecting additional or dropping existing items. More importantly, some items arguably belong in more than one subsector. #7 and 8 depend on adequate water supplies. Reassigning one of them to "Water supply" would balance item numbers (three in each subsector), though in an arbitrary manner. On practical grounds, it might be more fruitful to reformulate item questions under "Supplies" (water, hygiene products), "Maintenance" (storage containers, functioning latrines and showers) and "Waste disposal" (garbage collection, latrine pit emptying). This would, however, run afoul of official WASH divisions. Drinking water takes priority over showers.

Two complications are in the way of attempts to form subsector measures from the item ratings:

- **Importance weights:** Agency analysts may want to attach importance weights to the items. The rationale for that is plausible. Not all problems are equally important or potentially life-threatening. Drinking water takes priority over showers. Formally, however, importance weights on ordinal variables imply one of two things: 1. all ratings under the same items are multiplied by a constant – an operation not permitted for comparisons with other ordinal variables, 2. the item data are expanded horizontally – think of copying the same column multiple times. This creates data management as well as interpretation issues (it second-guesses the raters *aka* key informants).
- **Non-commutativity:** How should the measure of subsector priority be formed? If we stick to the median as the way to average ratings, should we first take the row medians, then the column median of the row medians for the ratings under a given sub-sector? Or first down the columns, then across the row? Mathematically the median of medians is not commutative. In other words, the result may differ depending on the sequence taken.

A way out of both problems at the same time leads through transformations of the ratings to some interval-level construct, and hence the switch from medians to arithmetic means. See the sidebar about converting ratings to metric variables, on page 21 sqq.

Regardless, initially the boundaries between subsectors should be ignored, and the decision to head for subsector-specific measures should be based on the statistical association pattern of the items and its meaningful alignment, if any, with subsector definitions.

Data are rankings

As noted, ranking maps objects to relationships with other objects of the same domain. An ordinal scale, with *several levels* with external conceptual definitions, is *not* used. Rather, *one common dimension* is needed in which objects are compared. The relationship is formed by a subjective order (e.g., preferences, where the ranking agent likes A better than B, and B better than C, etc.) or by a common metric (e.g., cities ranked by population size). The ensuing ranks are no longer statistically independent, unlike ratings.

Numbering

The numbering of ranks differs, depending on convention.

- In **track** ranking, the lowest value is ranked 1, and there is no correction for ties. That is, the track rank is 1 + the number of values that are lower (for the subjective variety, read: "1 + the number of items deemed less important in the common dimension).
- In **field** ranking, the highest value is ranked 1, and there is no correction for ties. That is, the field rank is 1 + the number of values that are higher.
- In **unique** ranking, the values are ranked 1, 2, .. N (N is the number of objects to rank). Values and ties are broken arbitrarily. E.g., two values that are tied for second are ranked 2 and 3.
- In **average** ranking, in the *ascending* type, the ranks are as in unique ranking, but with the tie ranks as the arithmetic means of the corresponding unique ranks. The *descending* variety is calculated as N + 1 – ascending average rank.

Table 12: Ranking modes

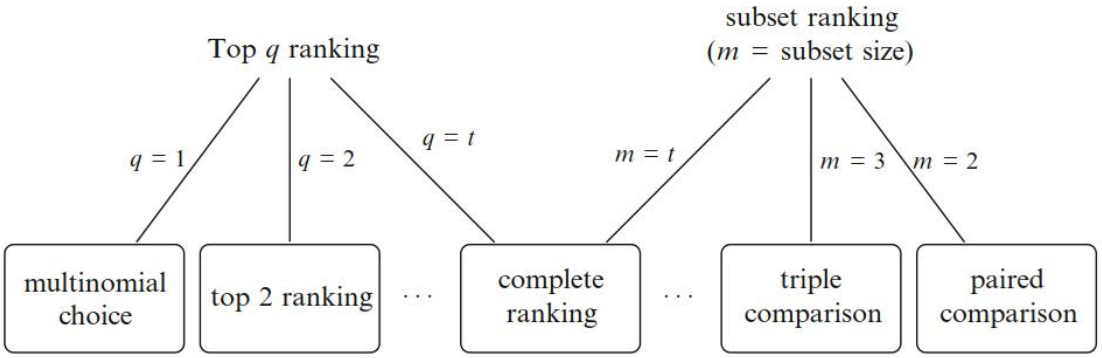
Objects		Ranking modes				
ID	Ranking basis	Track	Field	Unique	Average, ascending	Average, descending
I	1	1	8	1	1	8
II	5	2	5	2	3	6
III	5	2	5	3	3	6
IV	5	2	5	4	3	6
V	10	5	4	5	5	4
VI	12	6	2	6	6.5	2.5
VII	12	6	2	7	6.5	2.5
VIII	15	8	1	8	8	1

The Excel function RANK.AVG supplies the two flavors of average ranking. The function RANK.EQ with option "ascending" creates a track ranking, with option "descending" a field ranking.

Objects to rank

Ranking strategies are further distinguished by the completeness of the ranking operation and the number of objects that are ranked in a basic operation. The two major classes of ranking are top-*q* ranking and subset ranking. The diagram by Alvo and Philip (2014:2) facilitates understanding.

Figure 6: Classification of rankings



t = number of objects, q = number ranked, m = number ranked in lower-level (subset) ranking

In top-q rankings, only the q objects considered most important are ranked, the remainder are left unranked. The "partial Borda count" (see below), usually with three prioritized objects, is an instance thereof.

In subset ranking, in a first stage subsets are ranked. The types of subsequent operations vary widely. For example, ranking the winners from each subset and discarding the rest is common in situations when the objects to rank are divided among experts, with each subset evaluated by one expert (or by non-overlapping expert groups). Only the top one or two choices from each subset are communicated for the second-stage ranking (typically done with job applications in a firm).

[Sidebar:] Multi-stage ranking

Ranking should be used only when the respondents can cognitively handle all the objects at the same time. If the number of objects exceeds their memory and decisional capacity, the ensuing ranks will be incomplete or unreliable. If a complete ranking is desired, one possibility is to break the set of objects into subsets. This leads to multi-stage ranking. In situations like the Rohingya refugee re-surveys, which had key informants rate 17 needs areas, if ranking is desired, one may be tempted to pursue a multi-stage approach. The alternative would be multiple paired comparisons; but this might be unviable or likely to produce inconsistencies (A > B, B > C, C > A).

In a first stage, the respondents rank the objects within each subset. In further stages, the k-highest ranked members of each subset of the preceding stage are regrouped into one or more sets, and the respondents rank the objects in these. Realistically, in verbal elicitation, k will be 1 or 2.

Unless further questions probe the relationships among the lesser ranked objects of the earlier stages, this procedure will produce a semi-order (Wikipedia 2012), not a complete and coherent one. This may be acceptable if it reliably yields top-q ranked objects, say, the top three. There are reasons to doubt that it will.

Multi-stage ranking is path-dependent

The multi-stage ranking process is not associative, i.e., the final ranking differs depending on how the initial subsets were composed, and on how their elements were ranked. If the union of first-stage subsets includes all objects AND if the ranking basis is metric, then the final "winner among winners" should always be the same. This is regardless of subset formation and of the *q* in top-*q* ranking within the subsets. However, this does not apply to the 2nd, 3rd, etc.-ranked winners. And it does not apply to rankings from a subjective preference order that is not strictly transitive.

Suppose six objects (e.g., needs areas) {A, B, C, D, E, F} are to be ranked, and the respondent has strict preferences A > B > .. > F. Presenting A – C in a first subset, and D – F in the second will reveal the orders A > B > C and D > E > F.

How to rank in the second stage? Simply comparing {A, D} and {B, E} will reveal A to dominate all other objects, i.e. A is the undoubted winner of winners. However, it will not establish the final second-ranked object. Both B and D remain candidates. The alternative is to ask the respondent to rank {A, B, D, E} simultaneously. This produces unambiguous first and second-rank winners (unless, of course, the respondent recants the first-stage choices A > B and/or D > E).

It does not yet identify the third-ranked object. E is a candidate because of the revealed chain A > D > E. It is, however, not the only choice; C reappears from the first-stage order A > B > C. Thus, {C, E} have to be compared as part of the second-stage elicitation, in order to achieve a clear top-3 ranking from subsets.

To make matters more complicated, if the respondent's preferences are not strictly ordered, the outcome may be susceptible to the initial way of forming the subsets, e.g. {{A, B, C} {D, E, F}}vs. {{A, B, F} {C, D, E}}.

Prone to confusion and fatigue

By this point, it should be clear that multi-stage ranking is not really suitable for ranking *n* objects, *n* > 6, in face-to-face interviews with verbal elicitation and, at most, a visual aid presenting the objects as readily understood symbols. Apart from respondent confusion or fatigue, there would also be challenges of navigation for the interviewers and during the data entry. It may work well in self-administered Web-based surveys, but this is beyond our purview.

Chances to obtain complete rankings by multi-stage rankings are thus not great. This conclusion is derived from the difficulties to extract ranks from within-subset comparisons and the regrouping of top-ranked objects at the following stage. In other words, they are inherent of every individual interviewer-respondent interaction and, as such, have nothing to do with subsequent aggregation challenges.

The short version is: *Don't use multi-stage ranking questions in needs assessments.*

In fairness, we note that there is a small, but growing literature on multi-stage ranking. Virtually all of it speaks to Big-Data automated Web-based recommender or document retrieval systems (Culpepper, Clarke et al. 2016, Qu, Meng et al. 2016). Interestingly, a good part of the contributions discuss the trade-offs between what they call "effectiveness" (extensive and robust rankings, quality in the eyes of users or advertisers) and "efficiency"

(speed, cost). It is the same in comparatively small-sample surveys and assessments when translated to the conflict between completeness and elicitation effort.

A better way?

The better, if still daunting, alternative is to go for *partial* rankings from the beginning. This could all objects in the set (e.g., the 17 needs areas) from among which the three highest ranked are elicited; these preferences will eventually be aggregated into some priority measure (such as the mean Borda count).

Better even, in order to make the task more manageable, assessment designers may aim at several measures, each grouping needs areas of a specific response phase or character. One might, for example, place short-term physical needs in a first group. A second group would embrace protection-related sphere. Areas with long-term consequences, including reconstruction and education, would be placed in a third. The priority measures that come out of these groupings will each have worth by itself. It is not be necessary or helpful trying to combine them into a super-construct of any kind.

Avoid false choices

In general, two items X and Y should be placed in different groups if comparing them as priorities would imply a false choice. Asking a key informant to rank food relief vs. protection from sexual violence within the same set options implies a trade-off. If, say, food relief is a higher priority, lessening protection up to a certain point would *seem* to be defensible, in order to boost food deliveries. In real life, this trade-off is made neither in the moral and physical economy of the affected group, nor in the coordination among relief agencies.

Ratings as a training ground for rankings

There is another reason why rankings by separate item groups may work better. This has to do with ratings. Rating questions can (and often do) precede ranking questions. Rating questions are each specific of one needs area. Statistically, they are independent, unlike ranking questions. Cognitively, they are interdependent, as studies in context effects (particularly question order effects) in survey methodology show. Respondents will seek for common underlying concepts, "what are these questions all about?"

Assume, as above, that the questions come in separate groups, each group relatively small (4 – 8 needs areas). When the interviewer asks the rating questions from a given group in short succession, they function as a kind of cognitive training area[14]. By the time he/she elicits priorities in a ranking question, the respondent should understand what this is all about (e.g., protection-related), and should recall all or most of the items. Using the "rating first, ranking next" approach with separate and independent substantive item groups keeps the cognitive burdens lighter. This, indeed, is a form of subsetting, but it is

[14] This is a steep claim, and it is unclear which studies from cognitive science would support it for the specific context we describe here. Studies about the psychology of question-answer sequences (e.g., Graesser, Baggett et al. 1996) are mostly set in the teacher-student (and computer – user) context, in which students are expected to ask questions back in response to those asked by teachers and computer-based surveys. The enumerator – key informant relationship in needs assessments is more one-sided ("extractive" in pejorative terms), but it is likely that informants in many situations ask for clarifications about the questions they are being asked, and that enumerators encourage such requests informally.

not multi-stage. There is no necessary second stage to combine the measures resulting from the different item groups.

Column ranks vs. row ranks

In a data matrix, ranks can be established by row, or by column, or on the entirety of the relevant data elements. This figure depicts an artificial data set with six objects on which four ratio-level indicators were measured. For better orientation, the values were chosen such that the column rankings are identical for all indicators, and the row rankings are identical for all objects. The panel at the lower right gives the (average, ascending) ranks of the 6 * 4 = 24 cell values. These were directly calculated off the upper left panel. There is no way to predict them from the combinations of row and column ranks.

Table 13: Row, column and element ranks

Object	Indic1	Indic2	Indic3	Indic4
I	6	13	26	34
II	13	29	36	43
III	26	34	40	50
IV	31	49	57	60
V	40	52	65	73
VI	55	60	76	88

Object	RowRank1	RowRank2	RowRank3	RowRank4
I	1	2	3	4
II	1	2	3	4
III	1	2	3	4
IV	1	2	3	4
V	1	2	3	4
VI	1	2	3	4

Object	ColRank1	ColRank2	ColRank3	ColRank4
I	1	1	1	1
II	2	2	2	2
III	3	3	3	3
IV	4	4	4	4
V	5	5	5	5
VI	6	6	6	6

Object	All1	All2	All3	All4
I	1	2.5	4.5	8.5
II	2.5	6	10	13
III	4.5	8.5	11.5	15
IV	7	14	18	19.5
V	11.5	16	21	22
VI	17	19.5	23	24

For a set of pairs of objects and indicators/items, row and column rankings together are available only if

- The basis of ranking is numeric and one-dimensional (same basic unit, e.g. monetary), or
- There exists a complete ranking of all elements.

The first case is illustrated by the upper left panel. The second by the lower right; from these 24 ranks the same row and column rankings can be obtained.

Such a situation implies an observer who has complete knowledge either of the numeric basis or the element ranks. Almost by definition, needs assessments cannot find such an observer at the beginning, and probably not at any stage later either. Nobody has comparable figures for the damage to buildings, infrastructure and production facilities in hundreds of communities right away after an earthquake. The observers are many, scattered, with different degrees of information. Key informants each offering information, estimates and opinion about a local community belong here.

The function of ranks changes. They express an unobserved underlying measure of factual or anticipated humanitarian impacts – on the local object to which the individual observer speaks. Objects become observers, and indicators become items ranked by the observers' beliefs. The ranks are row-wise ranks. They are column-wise independent because every observer makes their own ranking of the items. They are dependent row-wise because ranking takes into account several items. Unless there are missing ranks, the row sum of ranks is identical for all observers.

This subtlety is justified by two contingencies:

- First, there may be several key informants in a community. The assessment protocol may call for separate interviews with officials, with a male as well as a female focus group. This produces three sets of ranks per community. Rather than averaging them by community, a two-way analysis by community and informant type may be the most instructive approach. Multi-observer multi-item ranking methods have been studied for situations as diverse as wine tasting (Cao and Stokes 2017) and language accessibility in the European Union (Ginsburgh, Moreno-Ternero et al. 2017). They are beyond the brief of this note. Here we consider only one observer per object.

- Second, the same data table may contain column-wise ranks from another procedure and from sources other than the local key informants. An example would be the ranking of the communities by Copeland's Method for which several variables capturing other aspects were used. The table may also include other community descriptors, including naturally ordinal ones such as the highest type of school present (none, pre-school, primary, secondary, etc.). In later analyses, the Copeland ranks and descriptors might become explanatory variables for the ranks that observers gave to certain items. If a variable reported the proportion of households living in makeshift shelters and damaged buildings, one would expect it to be correlated with the priority rank for shelter.

Table 14: Row-wise priority ranks, column-wise community ranks in the same table

Key informants speaking for	Items, e.g.				Ranking of communities on, e.g.	Other community descriptors, e.g.
	Water	Food	Shelter	Health care	Exposure to violence	Education
	Priority (row-wise rank)				Copeland rank (column-wise)	Highest school type present
Community 1	4	3	1	2	6	Primary
Community 2	3	4	1	2	3	Secondary
Community 3	4	3	2	1	4.5	Primary
Community 4	4	3	2	1	4.5	Primary
Community 5	4	1	2	3	1	None
Community 6	4	1	3	2	2	None

In the following, we discuss two methods – the Borda count, and the Plackett-Luce model.

Borda count

History and status

Like Copeland's Method, the Borda count came into being first as an election method (Borda 1781, Wikipedia 2011). It has since been adopted in Multi-Criteria Decision Making (MCDM), and the quality of its outcomes have been compared with numerous other methods (Lansdowne 1996). Disagreements about its relative merits persist (Risse 2005). Borda's contemporary and rival was the Comte de Condorcet, another early theorist of election methods (Wikipedia 2018a). To this day, Borda is considered the "father of the compensatory approach" while Condorcet spawned a progeny of non-compensatory ones (Munda and Nardo 2009:1514). Current research into Web-based recommender systems is further stimulating methodological debates surrounding Borda (Lestari, Adji et al. 2018). There have been increasing applications in environmental and humanitarian assessments as well (Burgman, Regan et al. 2014, Limbu, Wanyagi et al. 2015).

Basics and terminology

In the classic variant, N candidates vie for the same elected position. A voter gives N votes to his/her most favored candidate, $N - 1$ to the second choice, etc., and 1 to the least preferred. Ties are not allowed. Other variants number the preferences $N - 1$ to 0, but for complete voting this is irrelevant as long as it is done consistently. The candidate with the highest sum of votes is the winner.

Terminology is not uniform. Besides "votes", one finds "marks" and "points" for the number expressing the strength of the preference that a voter feels for a candidate. Regardless, the "Borda count" is the sum of votes, marks, or points that a candidate received. The "mean Borda count" is the count divided by the number of voters. The term "Borda score" is often used, although, it seems, not unambiguously. In rare places, it denotes the votes that voter X gives to candidate Y. Mostly, it means the count of all the votes that Y receives.

Needs assessments establish priority needs, possibly from a considerable number of needs areas. The needs areas are the items; the number of distinct areas is the analogue to the N candidates. The key informants ranking the needs are the equivalent of the voters; if there is only one informant by assessed unit (community, camp), the ranks will be stored in one record per unit. Key informants may have strong views on the most pressing needs of their communities, and weak and indifferent opinions on lesser needs. Assessment designs take that into consideration by asking only about the $n < N$ highest priorities. Typically, $n = 3$.

Partial voting

Borda counts with only $n < N$ elicited priorities are "modified" or "partial-voting Borda counts". The votes are recorded as n for the highest, $n - 1$ for the second highest, thus in the typical case as 3, 2, and 1. Unranked options are scored as zero votes. However, a zero

should be recorded only if the item was clearly presented to the informant as options. Else they should be recorded as blanks.

Ratio-level interpretation in elections

The Borda count is, as its name says, a count variable and as such formally a ratio-level statistic. It is natural and legitimate to say that candidate A received *x* times as many votes as B did. The mean Borda count, which takes into account missing values (for options not presented to particular voters), claims the same measurement level. However, in non-election uses, the interpretation should be more conservative.

Shortcomings

The question is whether the mean count truly reflects voters' preferences, and whether it is a valid measure of preference strength. Its validity is at stake because the Borda outcome is not robust to so-called "clones" and "irrelevant alternatives" (Wikipedia 2014a, Wikipedia 2014b). This is best explained by example.

> Suppose key informants in refugee camps rank the priority of relief in the sectors *food, water, sanitation, shelter* and *health care*. Suppose that the ranks are Borda-counted; the counts reveal food relief as the highest priority and drinking water provision as the second highest. In other refugee camps in the area too, people have higher unmet needs for food than for water. However, here the designers add "cash relief" to the options. People know that with more cash in hand they will be able to buy more food in the local market. This is not true of water; increases in the supply of water depend chiefly on relief agencies boring more wells. As a result, the majority of key informants in these camps rank priorities as: water > cash > food, etc. The priority order between food and water is reversed. This is so because "cash" stole more votes from "food" than from "water"[15].

When assignment designers anticipate clone and irrelevant-alternative effects, they should prune items that they believe will distort true priorities the most. To stay with the example, cash is fungible, and "cash relief" addresses needs in multiple sectors, to the extent of the local market response (or the receptivity of relief personnel and officials for bribes). Put differently, cash relief is in a "means-to-ends" rapport with needs for food, shelter, clothing and many more. While deserving of its own place in assessment instruments, cash relief should not be an item in the ranking of priority needs.

Advantages

Despite those shortcomings, the Borda count remains a useful method for establishing priority needs areas. Its major advantage is its simplicity. This works at several levels. Key informants understand requests to rank things, problems, needs, etc. in response to questions of the *"Of these* [previously discussed, read out again from a list, items], *which is the most important? And which the second most important?"*, etc., kind. The response can be recorded in a table in the questionnaire, with a row for each item, and check boxes

[15] This fictitious illustration mixes the effects of clones and irrelevant alternatives. For separate demonstrations of how Borda violates the axioms of independence of irrelevant alternatives and of clones, see the two Wikipedia articles, op.cit.

for 1st, 2nd, 3rd priority, and one for "was not given as an option". At data entry, the priorities can be recorded in polytomous or indicator form (Benini 2013a:23 sqq.); Excel functions exist to calculate the Borda points for each respondent and ranked item (Munsch and Benini 2013: worksheet "Text02_Priorities"). In the analysis, the calculation of the mean Borda counts (the scores of interest) is trivial.

Interval-level interpretation in needs assessments

Undoubtedly, the Borda count delivers a ranking of items. Technically, the ranking arises from the order of the column means. In the fictitious example shown in the next table, key informants in ten camps indicated first, second and third priorities from among six needs areas presented as options. In the left panel, the first priority option is scored 3, the second 2, the third 1, and the unranked 0, as usual. There are no missings.

Need #4 is the clear first priority. #3 is the least urgently felt need. One could be tempted to say that #4 is 2.1 / 0.3 = 7 times as urgent as #3. This, literally, would be the ratio-level interpretation.

Table 15: Borda count example – Two scorings for partial rankings

Key informant for camp #	Need #1	Need #2	Need #3	Need #4	Need #5	Need #6
1	3	2	1	0	0	0
2	3	2	1	0	0	0
3	3	2	1	0	0	0
4	0	0	0	3	2	1
5	0	0	0	3	2	1
6	0	0	0	3	2	1
7	0	0	0	3	2	1
8	0	0	0	3	2	1
9	0	0	0	3	2	1
10	0	0	0	3	2	1
Mean Borda count	0.9	0.6	0.3	2.1	1.4	0.7

Key informant for camp #	Need #1	Need #2	Need #3	Need #4	Need #5	Need #6
1	6	5	4	*2*	*2*	*2*
2	6	5	4	*2*	*2*	*2*
3	6	5	4	*2*	*2*	*2*
4	*2*	*2*	*2*	6	5	4
5	*2*	*2*	*2*	6	5	4
6	*2*	*2*	*2*	6	5	4
7	*2*	*2*	*2*	6	5	4
8	*2*	*2*	*2*	6	5	4
9	*2*	*2*	*2*	6	5	4
10	*2*	*2*	*2*	6	5	4
Mean Borda count	3.2	2.9	2.6	4.8	4.1	3.4

However, this is not the only rational way to score the key informants' priority choices. The zero value expresses our almost complete ignorance about the relative importance of the unranked options – complete but for the fact that they are less important than the ranked ones in the individual informants' beliefs. Moreover, the zero-scoring of the unranked option reflects the frequent experience of assessment teams that lower-priority needs are volatile and dimly articulated.

Assume for a moment that the key informants entertained clearly distinct priorities for all six options, but the interviewers recorded only the first three. In this case, it would be reasonable to score these 6, 5 and 4, and to assume that the informants did rank the three lesser priorities without much difficulty. For these, by the Borda rules, the scores would be 3, 2, 1. We ignore their actual distribution. We do know, however, the expectation – the mean of these ranks, which is (3 + 2 +1) / 3 = 2. For the general case of N options and $N - n$ unrecorded ones, $((N - n + 1)\ (N - n\ ) / 2) / (N - n) = (N - n + 1) / 2$. This approach is exemplified in the right panel.

Compared to the left panel, the ratio of the #4 Borda count to #3 dropped from 2.1 / 0. 3 = 7.0 to 4.8 / 2.6 = 1.8. The intervals have stayed the same within the two subgroups, #1 - #3, and #4 - #6. They have changed between the subgroups. This is so because the observed ranks were lifted by 3 points, the unobserved ones by 2 only. The new scoring causes one rank reversal – between #1 and # 6, from 0.9 > 0.7 to 3.2 < 3.4. The change is mild; it concerns two sectors of relatively low priority; both differences are well within the measurement error typical of humanitarian needs assessments.

An interval-level measure

The lesson to take from this thought experience is twofold. The strong conclusion is that *the ratio-level interpretation of the partial Borda count is inadmissible*. The weaker one is that there is a case for the interval-level interpretation. In other words, we can evaluate the distribution of the mean Borda count and make comments like, for a notional example: *"In this ranking, the Borda counts of the first and second priorities are far apart from those of the rest; the spread among these is smaller than their distance from the second. Therefore, the first and second priorities are likely to be genuinely higher."*

The following graph takes the argument for interval-level interpretation further. Hypothetical scenarios are shown for three Borda count results. In each, key informants ranked six needs. The results from the assessment sample are marked A to F from highest to lowest Borda counts. The real-life meanings of the letter symbols are immaterial; they need not be the same from one scenario to the next. Since first priorities are scored as three points, the theoretical range of the mean count is [0, 3]. Every informant assigns 3 + 2 + 1 = 6 points; thus the mean counts of all ranked areas, absent missings, sum to 6.

In this example we have with six needs areas, A – F. The mean counts under completely random choices would closely fluctuate around 6 / 6 = 1. This is marked with the vertical reference line off the x-axis at 1.

Figure 7: Interval-level interpretation of the Borda count

In every one of the three cases, we have a complete numerical ranking. If we use the Borda counts as ordinal information only, we are bound to consider A the first, B the second, etc., and F the last priority everywhere. However, in the first case, this conclusion may be spurious. The means differ little, compared to the range. Needs area A may have truly higher priority than F, less likely higher than E, etc. The difference between A and B is likely insignificant. For B, the probability that it has genuinely higher priority than F, E, D, C is even less. And what is "likely" in this constellation? Thus one might conclude that the basis of prioritization is insufficient; the differences, when seen as distance on an interval scale, are not big enough.

This is different for Cases 2 and 3. In #2, A and B distinguish themselves clearly from the rest. In #3, the distances ED and CB are much larger than FE, DC or BA. Thus we are inclined to assume areas A and B to be *comparatively* high-priority, C and D mid-level, and E and F low. The italics are important; we compare ranks, not ratings. In this usage of the Borda count, there is no substantive zero point.

An example with real data

Earlier in this note, we used a segment of the data from the Rohingya refugee site assessment, Round 9, conducted in February or March 2018. Round 11 was conducted in May and became available as a cleaned dataset in July. We use needs prioritization measures from this interim version in order to illustrate the interval-level Borda count interpretation. Round 11 is worth a fresh look because the list of needs area options excludes "cash". This option distorted the picture in Round 9.

The number of camp points, by Round 11, had risen to 1,998. In almost all of them (1,989), interviewers effectively elicited from key informants the first three priorities from among

thirteen needs areas, plus “Other needs”. Item-missings are virtually inexistent (one “Don’t know” for first priority). This table is sorted descendingly by the number of mentions as first priority; to the right the Borda count and the mean Borda count are shown. Note that there are several reversals when we compare the ranks by priority-1 frequencies vs. by mean counts. The most significant is between food assistance and water. The Borda count for water is bumped up by the significantly higher number that it received for second priority mentions.

Table 16: NPM Round 11 - Priorities by needs area, Borda counts - overview

Needs area	Priority 1	Priority 2	Priority 3	Total	Borda count	Mean Borda count
Cooking fuel and firewood	625	602	523	1,750	3,602	1.81
Food assistance	543	136	137	816	2,038	1.02
Water	506	434	164	1,104	2,550	1.28
Shelter	91	183	148	422	787	0.40
Job opportunities	59	176	369	604	898	0.45
Education for children	37	119	100	256	449	0.23
Safety and security	34	53	142	229	350	0.18
Sanitation	30	120	114	264	444	0.22
Healthcare	28	75	112	215	346	0.17
NFIs	21	60	112	193	295	0.15
Psychosocial support	2	10	17	29	43	0.02
Skills development	0	4	5	9	13	0.01
Mobility	0	1	4	5	6	0.00
Other	13	17	43	73	116	0.06
Do not Know	1	0	0	1		
Total	1,990	1,990	1,990	5,970		6.00

By comparing mean Borda counts, we get the following natural priority groupings. Within groups, areas are further listed descendingly by Borda count:

1. Fuel
2. Water, food
3. Jobs, shelter
4. Education, sanitation, safety, health care, non-food items
5. PSS, skills development, mobility

One suspects that many key informants did not understand the options that eventually landed in group #5. Regardless, the key finding is that three needs areas are dominant. Within these, fuel and firewood – with which to cook food - outranked food itself. All that had been known for quite some time. Less trivially, compared to the top tier, needs in the shelter, safety and health care areas overall were less pressing; actors in these sectors must have succeeded in filling them to a significant degree (the data were collected before the peak of the 2018 monsoon).

Calculation

The priority choices were initially entered in polytomous format[16], as seen in the six first records of the data table.

Table 17: Sample records of the three priority variables

Priority 1	Priority 2	Priority 3
Food assistance	Cooking fuel and firewood	Job opportunities
Cooking fuel and firewood	Job opportunities	Education for children
Food assistance	Cooking fuel and firewood	Healthcare
Food assistance	Sanitation	Education for children
Food assistance	Cooking fuel and firewood	Healthcare
Cooking fuel and firewood	Healthcare	Shelter

The calculation of the priority frequencies and subsequently of the Borda counts off the polytomous variables is relatively straightforward. It involves three Pivot tables, manual merging of copies, and sorting. Once this is in place, the Borda count results from a trivial weighted aggregation, and the mean counts by the division by the number of non-missing values[17].

That procedure is efficient if one wants to limit the analysis to the global overview. If population-weighting or breakdowns by other variables are wanted, it may be more productive to calculate the Borda scores for each needs area to the right side of the data table, create named ranges for these column variables, then call the names in the analysis table. The demo workbook shows the detailed mechanics. The essential tools for the Borda scores are mixed cell references and the ISERROR and MATCH functions. If one has to reckon with case-wise missings (all three priorities missing), COUNTBLANK will quicken the work and avoid inflating the denominator for the mean Borda count. Here is an example; it should be easy to adjust the formulas to other data situations.

[16] "Polytomous": having more than two unordered categories, which may appear as text variables with spellings controlled by drop-down menus. The chief alternative to polytomous is the indicator format.

[17] The NPM field analyst created also priority variables in indicator format, i.e. 17 indicators off each of the three polytomous variables = 51 new variables. The Borda scores (the values for every camp point and every needs area) can then be placed in yet 17 more variables. But this intermediate step is not necessary.

Table 18: Extracting Borda scores from polytomous priority variables

	27	28	29	30
1	**Priority 1**	**Priority 2**	**Priority 3**	**Cooking fuel and firewood**
2	Food assistance	Cooking fuel and firewood	Job opportunities	2
3	Cooking fuel and firewood	Job opportunities	Education for children	3
4	Food assistance	Cooking fuel and firewood	Healthcare	2
5	Food assistance	Sanitation	Education for children	0
6	Food assistance	Cooking fuel and firewood	Healthcare	2
7	Cooking fuel and firewood	Healthcare	Shelter	=IF(COUNTBLANK(RC27:RC29) =3,"",IF(ISERROR(MATCH(R1C, RC27:RC29,0))=FALSE,4-MATCH(R1C,RC27:RC29,0),0))

The formula shown in R7C30 is identical up and down column 30. Above, in R2C30:R6C30, are the results for the first five records. Note that it is zero in row 5; fuel and firewood is not one of the three top priorities in that record; therefore, by the scoring rules of partial-voting Borda, zero votes are given. The spelling of the column header must agree exactly with that of the corresponding value in columns 27 – 29. How the formula works is explained in the Excel companion workbook to Benini (2013a), sheet "Text02_Priorities"[18].

From scores to mean counts

In order to succinctly and safely access the column-wise Borda scores for every needs area, we name their ranges by their column headers. A subtlety of named ranges is that spaces are not tolerated. When names are created by the menu facility Formulas – Create from selection – Top row, Excel automatically replaces spaces by underscores. Once this is done, the function INDIRECT comes in handy for the efficient Borda count calculations (see the formulas in the next para).

Population weighting

A check on the sensitivity of the counts to the highly variable camp point populations may be appropriate. Population size may be associated with unobserved factors that affect people's priorities. If the differences between weighted and unweighted counts are large, which version to prefer may be difficult to decide. Perhaps the most important factors are globally known from other sources. For example, camp points with small populations may abound in remote places that are logistically less accessible and thus less well covered by certain sector agencies. It is also possible that key informant bias or interviewer forgery are correlated with the size of the populations.

[18] Available at http://aldo-benini.org/Level2/HumanitData/Acaps_How_to_approach_a_dataset_Part_2_Analysis.xlsm .

Table 19: Unweighted and population-weighted Borda counts

	1	2	3	4	5	6
1	Needs areas (underscores replace spaces)	Borda count	Mean Borda count	Pop-weighted Borda count	Pop-weighted mean Borda count	Difference weighted - unweighted
2	Cooking_fuel_and_firewood	3,602	1.81	3,687.2	1.85	0.04
3	Water	2,550	1.28	2,583.9	1.29	0.01
4	Food_assistance	2,038	1.02	2,043.5	1.02	0.00
5	Job_opportunities	898	0.45	911.0	0.46	0.00
6	Shelter	787	0.40	753.1	0.38	-0.02
7	Education_for_children	449	0.23	423.3	0.21	-0.01
8	Sanitation	444	0.22	420.7	0.21	-0.01
9	Safety_and_security	350	0.18	397.3	0.20	0.02
10	Healthcare	346	0.17	295.0	0.15	-0.03
11	NFIs	295	0.15	288.6	0.14	0.00
12	Psychosocial_support	43	0.02	45.4	0.02	0.00
13	Skills_development	13	0.01	13.6	0.01	0.00
14	Mobility	6	0.00	12.4	0.01	0.00
15	Other	116	0.06	109.9	0.05	0.00
16			**6.00**		**6.00**	

with the column ranges in the data table named as in column 1 of this table and as "Population" - hence these formulas:

Column	Formula
C2	=SUM(INDIRECT(RC1))
C3	=AVERAGE(INDIRECT(RC1))
C4	=SUMPRODUCT(INDIRECT(RC1), Population) / AVERAGE(Population)
C5	=SUMPRODUCT(INDIRECT(RC1), Population) / SUM(Population)
C6	=RC[-1]-RC[-3]

The differences between the weighted and unweighted Borda counts are small. They are relatively larger for the health care area. It may be worth testing for differences, by walking time to the nearest clinic, in average camp point population as well the mean Borda counts.

Differences by groups or attributes

We explore the conjecture that key informants speaking for camp points farther removed from a static health clinic tended to express a higher priority for the health care area.

Table 20: Priority of health care needs in response to distance from nearest facility

	1	2	3	4	5	6
1	**Walking distance**	**Camp points**	**Population affected**	**Average population**	**Mean Borda count**	**Population-weighted**
2	15 mins walk or less	1,073	511,806	477.0	0.11	0.11
3	16 to 30 mins walk	743	339,828	457.4	0.23	0.19
4	31 mins to 1 hour walk	117	44,444	379.9	0.43	0.28
5	More than 1 hour walk	20	6,965	348.3	0.10	0.05
6	No access to static health facility	37	15,520	419.5	0.16	0.07
7	**Total**	**1,990**	**918,563**	**461.6**	**0.17**	**0.15**

Calculated with the following formulas conditional on walking distances:

Camp points	=COUNTIF(Walking_distance,RC1)
Population affected	=SUMIF(Walking_distance,RC1, Population)
Average population	=RC[-1] / RC[-2]
Mean Borda count	=AVERAGEIF(Walking_distance, RC[-4], Healthcare)
Population-weighted	{=SUMPRODUCT(IF(Walking_distance = RC1, Healthcare), IF(Walking_distance = RC1, Population)) / SUM(IF(Walking_distance=RC1, Population))}

where "Walking_distance", "Population", and "Healthcare" are the names of data ranges. The formulas for the population-weighted Borda count are entered as single-cell array formulas, i.e. at first in R2C6, by pressing Shift-Ctrl-Enter. Then copy and paste to R3C6:R6C6. For the total (R7C6), use the unconditional expression =SUMPRODUCT(Healthcare, Population) / SUM(Population).

We note that the mean Borda count for health care needs, while overall low, does increase from camp points within 15 minutes to those within 31-60 minutes' walking time to the nearest static health facility. Interestingly, this is not the case for those who walk for more than one hour, or those that have no access at all. While these two sub-populations are concentrated in only 57 of the 1,990 camp points with data, one might suppose that their health care priorities were relegated by even higher needs in other sectors. Only a complete profile of mean Borda counts over all sectors and walking distance strata could possibly illuminate this hypothesis.

When we weight the mean Borda counts by the population size of the camp points, the effect of the walking distance further diminishes. Within each walking size stratum, more populous camp points tend to assign even lesser priority to health care, relative to other needs.

Two points can be made to conclude this sub-section:

Pro and contra population weighting: Comparisons between unweighted and weighted statistics are a useful *cautionary* exercise in many situations. Once must not forget, though, that the record of even a large camp point represents *only one* observation, i.e. just the

information given by one key informant, or a group of informants interviewed together. Even if we find that population weighting changes a result considerably, this tells us nothing yet about the specific mechanism that moderates some effect of interest apace with population size. Moreover, the effect may be spurious, i.e. the differences are driven by key informant bias aligned with population size.

Highly adaptable Excel functions: The type of array formula used in the population weighting by walking distance stratum above is adaptable to many other purposes, in weighted and unweighted situations. For example, had we wanted the median populations by stratum, then

{=MEDIAN(IF(Walking_distance = RC1, Population))}

would produce it. The general rules are:

1. The name of the main function is outside.
2. The IF-function is inside, with the criteria range = specific criterion value, first.
3. The range to which the function is applied follows, separated by a comma only[19].
4. Enter as array formula in one cell, Ctrl+Shift+Enter
5. Thence copy-paste to the remaining cells of the target area.

For complex expressions, the *IF-condition* will have to be *repeated* in every argument:

{=SUMPRODUCT(*IF(Walking_distance* = RC6, Healthcare), *IF(Walking_distance* = *RC6*, Population)) / SUM(*IF(Walking_distance=RC6*, Population))}

Subsector measures from sector-wide Borda counts

The situation outlined for ratings, under *"Items attached to subsectors of a given sector"* on page 37, may occur for ranking variables, too. Informants are asked to rank the three most important problems sampled from a sector's problem list. The problems are presented together, perhaps after the informants had rated each of them on a common severity scale. Cognitive overload is a concern; yet formally the mean Borda counts from the ranks of each problem are unproblematic. The analysts, however, wish to use the rankings to determine the relative priorities of the subsectors, such as Water, Sanitation and Hygiene within the WASH sector. They take the grand mean of the Borda-coded ranks for the all the problems under a given subsector.

This is problematic, for two reasons:

- First, as noted for the ratings case, the strict mapping of problems on administratively defined sub-sectors may be misleading. This is a categorical challenge.
- Second, a statistical challenge arises when the number of problems from which the rankings were elicited differs across subsectors. The intent to compare subsectors

[19] Placing the parentheses at {=MEDIAN(IF(Walking_distance = RC1), Population)} causes an error.

implies a null hypothesis that the problems in them are equally severe. The Borda measurements are taken in order to tell whether those in some particular subsector(s) are more severe – the alternative hypothesis. However, from the universe of possible problems, the assessment designers consider only some (those known from local observation or from sector doctrine); these become their sample (= item samples; the sample of respondents is the same for all subsectors). With unequal sizes, the probability that a random choice (e.g., for first rank) falls on a problem within subsector A (but not B or C, etc.) now depends on the number of problems subsumed under A. Thus the Borda scores have to be reweighted by the inverse of these probabilities, and the mean Borda counts from the reweighted scores have to be rescaled so that they sum to 6 (for top-3 ranked).

Not only are such calculations onerous (and intransparent to anyone not given a procedural logbook copy), it is even harder to understand the rationale for the exercise. If the problems under one subsector clearly dominate, this can be seen quickly from first visual inspection of the problem-wise mean Borda counts. Just as likely, the most important problems occur under different subsectors, side by side with less important ones. If so, the subsector means will, by definition, be smaller than the means of the top problems within them, and will thus be less informative.

We should be open to the possibility that analysts may come up with a stronger rationale for aggregating problem-level Borda counts to subsector priority measures. Yet, just by contemplating the interdependencies of problems and remedies that cross the WASH subsectors, one is inclined to think that unlikely.

The Borda count – Summary remarks

In summarizing, we want to reiterate one point: Formally, the mean Borda count is a ratio-level statistic. Substantively, however, the interpretation should be more conservative, treating the mean count as an interval-level measure. Ratio-like comparisons should be avoided. This should not prevent this method from correctly identifying top priorities when they do exist in the affected population.

The Borda count makes a number of assumptions. These are more productively discussed in contrast to the Plankett-Luce method, in the next section.

The Plackett-Luce model

Motivation

The interval-level measure of priorities that the Borda count delivers may not be satisfactory in all situations. Doubts arise in particular about the urgency of the lesser priorities. In the NPM Round 11 illustration, the drop from the three top-ranked priority needs areas (Fuel and firewood, water, food) to the next two (job opportunities and shelter) is particularly dramatic. But just how important are these latter needs still? The fact that overall they are distinctly more highly ranked than the following areas (education, sanitation, etc.) does not adequately answer that question.

The Plackett-Luce model supplies a stronger measure. It is ratio level. Moreover, it is based on more intuitive behavioral assumptions than the Borda count. Borda implicitly makes several strong assumptions. Respondents are aware that they exercise voting. They understand that they have exactly X points to distribute as votes. They have the entire set of options in mind. Individually, the strengths of their preferences are fairly expressed by integer quantities (3, 2, 1). Non-voted items do not contribute (i.e., are scored zero).

The Plackett-Luce model is different. Luce (Luce 1959, 1977, Luce 2008) supplied the individual choice theory underlying it. It assumes that actors, given a limited set of choice options, will select a particular option with a probability proportionate to its intrinsic worth. The marginal utility of the option, however, decreases with the already achieved degree of satisfaction. Thus, in refugee camps with scant water supplies, the increase in daily supplies from 4 to 5 liters is worth much more than from, say, 20 to 21 liters per person. The personal utilities of different options are not directly observable. They can be inferred, collectively, from stated or revealed preferences (the latter from market purchases or the use of flexible vouchers, etc.).

Implemented in R

Subsequently, Plackett contributed the statistical theory on estimation and inference (Plackett 1975). Hence the name Plackett-Luce model, which has become well established in the study of individual choice. Very recently, Turner et.al. made software available for estimates from partial rankings if every respondent ranks at least two options (Turner, van Etten et al. 2018). The R package "PlackettLuce" (Turner, Kosmidis et al. 2018) returns a ratio-level measure of the average worth of options in a population. In the humanitarian context, "worth" may be interpreted as "strength" of priority or preference[20]. The model has a solid probabilistic foundation and as such allows for testing for significant differences among options.

Coding and profiling in the PlackettLuce procedure

Stated priorities are coded differently in PlackettLuce from the Borda count. The first priority is coded 1, the second 2, etc. Unranked items are coded 0, as in the partial Borda. In two respects, Plackett-Luce is more flexible than Borda. In partial ranking of N items to choose from and n top ranks given, n is variable; individual rankers can rank any number > 1 of the given options, including full ranking; thus $N \geq n > 1$. Second, ties are allowed; a ranker may assign the same rank to several options.

[Sidebar:] Visualizing preference profiles with decision trees

In additional, the PlackettLuce package offers a feature that potentially is of great interest in humanitarian analysis. In combination with methods of random forests (optimal decision trees) (Wikipedia 2018c, Wikipedia 2018b), one of the procedures identifies and visualizes different preference profiles by ranker attributes (e.g., belonging to certain social groups, household characteristics, regions, etc.). Working with Borda counts, group profiles are

[20] We are not aware of humanitarian analyses using the model, with the tangential exception of Cardenas et.al., who experimented with Colombian public servants' preferences in allocating relief vouchers, including to refugees (Cardenas and Sethi 2010).

easy to make, but only descriptively for predefined groups. By contrast, PlackettLuce can identify groupings on the basis of several (categorical or continuous) variables that differentiate profiles optimally.

The procedure comes with data preparation requirements that appear complicated. For this note, therefore, we only present a borrowed illustration (Turner, Kosmidis et al. 2018 op.cit., 13-14). An NGO in Nicaragua had farmers each grow three experimental varieties of bean during one of the growing seasons, plus the standard local variety. At the end of the season, the farmers were asked to rank the four varieties. All in all, eleven varieties were grown and, in fours at a time, ranked for suitability. Three profiles were created by specific combinations of year (2015, 2016), planting season (one of three) and maximum night-time temperature (a continuous covariate). This figure reveals distinct preference patterns. Farmers in cooler seasonal environments clearly preferred varieties #7 and 9 for springtime (Pr) planting and felt milder preferences for, or disappointment with, all the others (Node 3). In the same climate zone, farmers experimenting in other planting seasons, expressed less well articulated preferences; varieties #8 and 9 did somewhat better (Node 4). Finally, farmers in warmer environments, regardless of year and season, were even less discerning in their preferences for any of the eleven varieties.

Figure 8: Plackett-Luce tree from an agricultural experiment

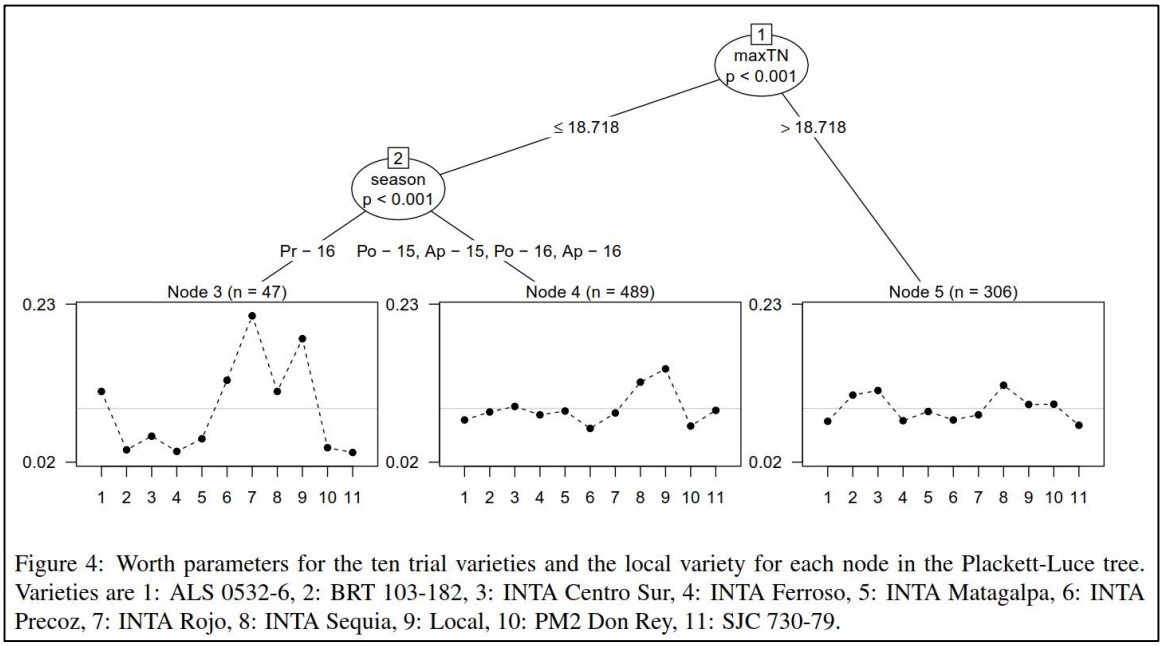


Figure 4: Worth parameters for the ten trial varieties and the local variety for each node in the Plackett-Luce tree. Varieties are 1: ALS 0532-6, 2: BRT 103-182, 3: INTA Centro Sur, 4: INTA Ferroso, 5: INTA Matagalpa, 6: INTA Precoz, 7: INTA Rojo, 8: INTA Sequia, 9: Local, 10: PM2 Don Rey, 11: SJC 730-79.

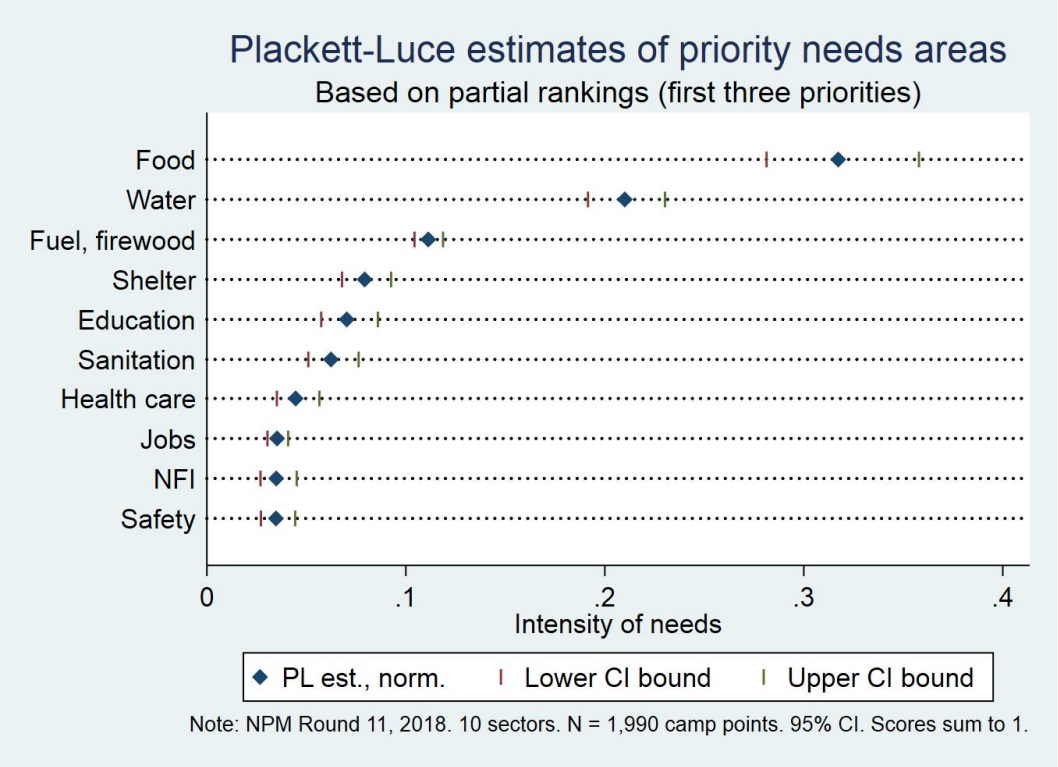
Apart from the attractive visual, one other important strength of this method is that the rankers are not required that all of them choose from the entire *N* options. Obviously, there must have been sufficient overlap across participants among the subsets of 4 varieties in each climate zone, year and season – else the results for all 11 varieties could not have been estimated conjointly. The overlap was produced by the experimental design. How sensitive the estimates are to variable levels of overlap is not discussed in this publication.

Results from the Bangladesh NPM Round 11 data

We ran the Plackett-Luce procedure on priority rankings for 10 of the 14 needs areas reported earlier. We excluded psychosocial support, skills development and mobility because they were rarely chosen as priorities. We excluded the residual option "Other" because of its mixed character.

The NPM site assessments in the Rohingya refugee camps are full enumerations; therefore, there is no sampling variance. However, there is model uncertainty, which is why confidence intervals make sense. The Plackett-Luce estimates of the true worth of items underlying the rankings are identified up to a scaling factor; in other words, they can be multiplied by any positive constant without disturbing the ratios between item values. We have normalized them such that they sum to one (and for comparison will do the same to the Borda counts). The figure shows estimates for the ten needs areas, sorted descendingly, together with their CI bounds.

Figure 9: Plackett-Luce priority estimates, with confidence intervals

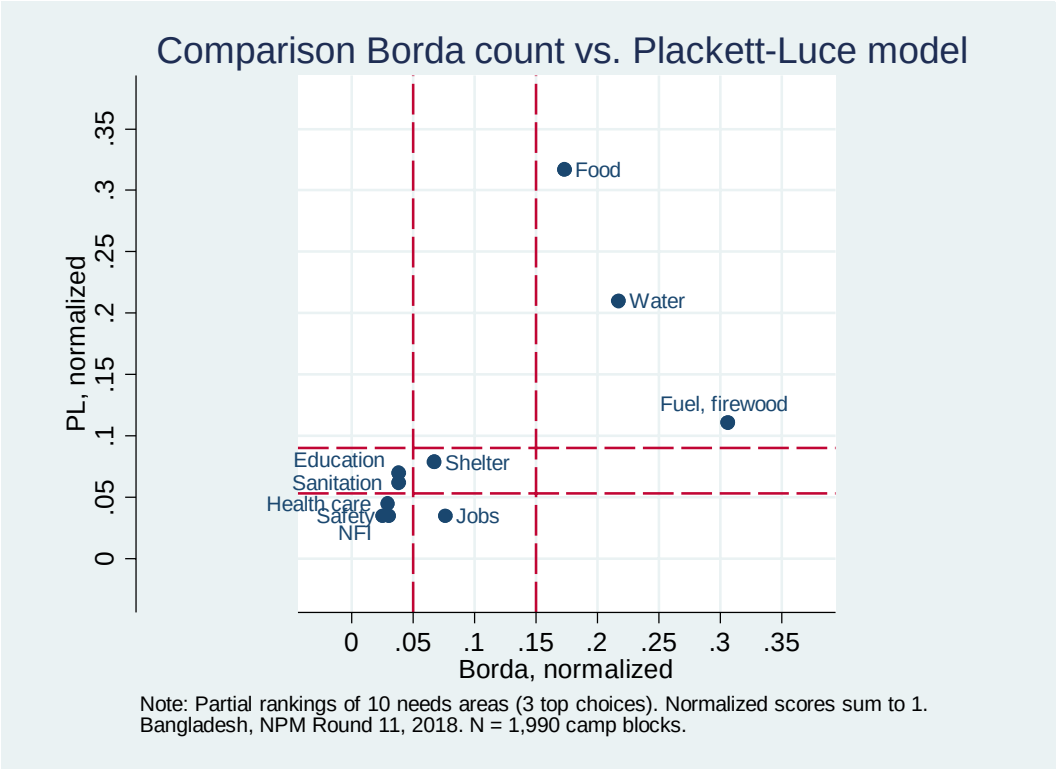


The priority of food relief is extremely strong. The top three needs areas – food, water, fuel-firewood – are clearly distinct among themselves and from the rest; their confidence intervals do not overlap. Starting with shelter, we look at a semi-order whereby shelter, education and sanitation are distinctly higher priorities than a (decreasing) number of lower-priority items. Health care, jobs, NFI and safety are indistinguishable in terms of priority.

Borda vs. Plackett-Luce

The following scatterplot compares the priority estimates under the two methods. Note again that the Borda count is interval-level; it has no meaningful zero point. The values may be shifted by adding any constant. Hence we can compare ranks and, with caution, distances, but not ratios. Nonetheless, the differences between the two methods are striking.

Figure 10: Borda vs. Plackett-Luce priority scores for ten needs areas



Note: Partial rankings of 10 needs areas (3 top choices). Normalized scores sum to 1.
Bangladesh, NPM Round 11, 2018. N = 1,990 camp blocks.

First, the agreements. Both methods place food, water and fuel/firewood in a distinct group of top priority needs. Both place shelter in a range that one may interpret as middling priorities. For both, health care, safety and NFI are comparatively low priority.

The most dramatic difference is the switching of places between food and fuel/firewood. Less distinctive are the contrasts in the middling zone. Following the Borda count method, jobs belong there. Plankett-Luce places sanitation and education higher up.

Why these differences? We focus on food vs. fuel/firewood, high priorities, on which one might expect better agreement between methods. The original priority rankings for them were:

Table 21: The three top needs areas, by relative rank frequencies

Needs area	Priority 1	Priority 2	Priority 3	Total
Cooking fuel and firewood	625	602	523	1,750
Food assistance	543	136	137	816
Water	506	434	164	1,104
Row-wise proportions:				
Cooking fuel and firewood	36%	34%	30%	100%
Food assistance	67%	17%	17%	100%
Water	46%	39%	15%	100%

Fuel/firewood gathered more choices at all three top ranks than any other needs area. Unsurprisingly, it has the highest Borda count. On the same measure, water outdid food assistance because of the much more numerous camp points that elected water as the second priority.

However, Plackett-Luce does not score observed ranks mechanically, as Borda does. Rather, it considers the tendency for items to be at the higher end of the observed ranks and then extrapolates that information into simulated rankings of the unobserved ranks. As we can see, within the observed priorities of each of the three needs areas, food assistance has a far higher share of Priority-1 votes than fuel/firewood and water. In camp points that ranked none of the three areas in the top three priorities, Plackett-Luce "believes" that among the unobserved priorities nos. 4 – 10 food assistance would likely have ranked higher than the other two if the key informants had ranked all ten. For the same reason, water is more likely to be higher up among priorities nos. 4 – 10 than fuel/firewood. In other words, Plackett-Luce uses the observed information to narrow down probabilities of the unobserved rankings[21].

The Borda count is indifferent to the unobserved rank orders. It is uninformative in this regard. All unobserved ranks receive a score of zero and go into the calculation as such.

The big question is whether the fundamental assumption of Plackett-Luce is realistic in humanitarian needs assessments. We doubt so more strongly where affected communities live in different environments – different on account of prior histories and current challenges and opportunities. Would the intrinsic values of the different types of relief and protection be the same for all, apart from strictly local idiosyncrasies? Or would we rather find situations in which the most urgent unmet needs are nearly the same for all communities, yet there is great diversity and volatility among the less pressing ones?

[21] Statisticians may find this explanation awkward or outright misleading, but it is implied in the treatment of the standard Plackett-Luce model by a rank-ordered logit model (Turner, Kosmidis et al. 2018 op.cit.: 2). Let respondent r choose item i. Let U_{ri} be the utility that r gains from i. Let α_i be the intrinsic value of item i in the population, and ε_{ri} the individual disturbance, i.i.d., such that $U_{ri} = \log(\alpha_i) + \varepsilon_{ri}$. Assume i = food, j = fuel. If ε_{ri} and ε_{rj} both << 0, neither may be one of the observed priorities for i but if $\alpha_i > \alpha_j$, it is more likely that i stands higher than j among r's unobserved priorities.

The Borda count's limitations are almost the opposite of Plackett-Luce's. Its priority profiles are purely descriptive – there are no confidence intervals. But the counts are perfectly additive. If we divide the NPM-covered camp points into two, and the Borda count for water in the first one thousand camp points is 1,000, and for the second it is 1,500, then the count for the total refugee area is 2,500. The additivity is one of the Borda count's properties that make it simple. Plackett-Luce lacks this property.

But Borda is too simplistic in its treatment of the unranked options. Suppose that among the 2,000 camp points, 200 made education a priority, and all of them ranked it first – no second or third priorities. The mean Borda count would work out as (3 * 200) / (6 * 2,000) = 0.05, a low value on a scale that ranges from 0 to 3. In Plackett-Luce, education would receive a higher score; the model takes into account that all the observed ranks for education are first ranks.

Overall, however, the result from the NPM Round 11 gives us the confidence that both methods will agree in the identification of a group of high priority needs.

References

Alvo, M. and L. Philip (2014). Statistical methods for ranking data, Springer.

Benini, A. (2013a). "How to Approach a Dataset : Part 3 - Analysis." Geneva, ACAPS. Retrieved 3 March 2016, from http://www.acaps.org/sites/acaps/files/resources/files/how_to_approach_a_dataset-part_3_analysis_march_2013.pdf.

Benini, A. (2013b). Severity and priority - Their measurement in rapid needs assessments Geneva, Assessment Capacity Project (ACAPS).

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

Benini, A. (2018) "Subjective Measures in Humanitarian Analysis [January 2018]." from http://aldo-benini.org/Level2/HumanitData/ACAPS_Subjective%20Measures__Jan_2018.pdf.

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

Burgman, M. A., H. M. Regan, L. A. Maguire, M. Colyvan, J. Justus, T. G. Martin and K. Rothley (2014). "Voting Systems for Environmental Decisions." Conservation Biology **28**(2): 322-332.

Cao, J. and L. Stokes (2017). "Comparison of Different Ranking Methods in Wine Tasting." Journal of Wine Economics **12**(2): 203-210.

Cardenas, J. C. and R. Sethi (2010). "Resource Allocation in Public Agencies: Experimental Evidence." Journal of Public Economic Theory **12**(4): 815-836.

Cliff, N. (1993). What is and isn't measurement. A Handbook for data analysis in the behavioral science. G. Keren and C. Lewis. New York, Lawrence Erlbaum Associates, Inc.**:** 59-94.

Culpepper, J. S., C. L. Clarke and J. Lin (2016). "Dynamic Trade-Off Prediction in Multi-Stage Retrieval Systems." arXiv preprint arXiv:1610.02502.

De Boeck, P., M. Wilson and G. S. Acton (2005). "A conceptual and psychometric framework for distinguishing categories and dimensions." Psychological Review **112**(1): 129-158.

Ginsburgh, V., J. D. Moreno-Ternero and S. Weber (2017). "Ranking languages in the European Union: Before and after Brexit." European Economic Review **93**: 139-151.

Graesser, A. C., W. Baggett and K. Williams (1996). "Question-driven Explanatory Reasoning." Applied Cognitive Psychology **10**(7): 17-31.

Greco, S., J. Figueira and M. Ehrgott, Eds. (2016). Multiple criteria decision analysis. State of the Art Surveys. New York, Springer.

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

Lättman, K., M. Friman and L. E. Olsson (2016). "Perceived accessibility of public transport as a potential indicator of social inclusion." Social Inclusion **4**(3): 36-45.

Lestari, S., T. B. Adji and A. E. Permanasari (2018). Performance Comparison of Rank Aggregation Using Borda and Copeland in Recommender System. 2018 International Workshop on Big Data and Information Security (IWBIS).

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

Luce, D. R. (1959). Individual Choice Behavior: A Theoretical Analysis. New York, Wiley.

Luce, R. D. (1977). "The choice axiom after twenty years." Journal of Mathematical Psychology **15**(3): 215-233.

Luce, R. D. (2008) "Luce's choice axiom." Scholarpedia, 3(12):8077. from http://www.scholarpedia.org/article/Luce%27s_choice_axiom.

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

Munsch, B. and A. Benini (2013). How to Approach a Dataset : Part 2 - Data Preparation (Excel file, macro-enabled) [Available from https://www.acaps.org/sites/acaps/files/resources/files/how_to_approach_a_dataset-part_2_data_preparation.xlsm]. Geneva
ACAPS.

Plackett, R. L. (1975). "The Analysis of Permutations." Journal of the Royal Statistical Society. Series C (Applied Statistics) **24**(2): 193-202.

Qu, J., X. Meng and H. You (2016). "Multi-stage ranking of emergency technology alternatives for water source pollution accidents using a fuzzy group decision making tool." Journal of Hazardous Materials **310**: 68-81.

Risse, M. (2005). "Why the count de Borda cannot beat the Marquis de Condorcet." Social Choice and Welfare **25**(1): 95.

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

Saisana, M. (2015). COIN tool: A do-it-yourself guide in Excel for constructing and assessing composite indicators. Ispra, Italy, European Commission, Joint Research Centre. Beta Version 2.0: November 2015.

Saisana, M. (2016). Step 6: Aggregation rules: Non-compensatory approaches. COIN 2016 - 14th JRC Annual Training on Composite Indicators & Scoreboards, 26-28/09/2016. Ispra (IT), The European Commission's science and knowledge service and Joint Research Centre.

Saisana, M. and A. Saltelli (2011). "Rankings and Ratings: Instructions for Use." Hague Journal on the Rule of Law **3**(2): 247-268.

Schröder, C. and S. Yitzhaki (2017). "Revisiting the evidence for cardinal treatment of ordinal variables." European Economic Review **92**: 337-358.

Turner, H., I. Kosmidis and D. Firth (2018). PlackettLuce: Plackett-Luce Models for Rankings. URL https://CRAN.R-project.org/package=PlackettLuce, R package version 0.2-3.

Turner, H. L., J. van Etten, D. Firth and I. Kosmidis. (2018). "Modelling rankings in R: the PlackettLuce package." Retrieved 10 December 2018, from https://arxiv.org/pdf/1810.12068.

van Herk, H. and M. van de Velden (2007). "Insight into the relative merits of rating and ranking in a cross-national context using three-way correspondence analysis." Food Quality and Preference **18**(8): 1096-1105.

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

Wikipedia. (2012). "Semiorder." Retrieved 18 April 2012, from http://en.wikipedia.org/wiki/Semiorder.

Wikipedia. (2014a). "Independence of clones criterion." Retrieved 25 April 2015, from http://en.wikipedia.org/wiki/Independence_of_clones_criterion.

Wikipedia. (2014b). "Independence of irrelevant alternatives." Retrieved 25 April 2015, from http://en.wikipedia.org/wiki/Independence_of_irrelevant_alternatives.

Wikipedia. (2016a). "Copeland's method." Retrieved 21 November 2018, from https://en.wikipedia.org/wiki/Copeland%27s_method.

Wikipedia. (2016b). "Ridit scoring." Retrieved 19 June 2018, from https://en.wikipedia.org/wiki/Ridit_scoring.

Wikipedia. (2018a). "Condorcet method." Retrieved 21 December 2018, from https://en.wikipedia.org/wiki/Condorcet_method.

Wikipedia. (2018b). "Decision tree learning." Retrieved 30 December 2018, from https://en.wikipedia.org/wiki/Decision_tree_learning.

Wikipedia. (2018c). "Random forest." Retrieved 30 December 2018, from https://en.wikipedia.org/wiki/Random_forest.

Wikipedia. (2019). "Softmax function." Retrieved 11 January 2019, from https://en.wikipedia.org/wiki/Softmax_function.

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.

Appendix

R-code for the Plackett-Luce model

```
# ========================================================================
# Subject:    ESTIMATION OF A PLACKETT-LUCE MODEL OF SECTORAL PRIORITIES
# Authors:    Aldo Benini, with help from Attilio Benini
# Date:       2019-01-03
# Data source:Bangladesh, Rohingya - NPM Site Assessment Round 11, 2018
# ========================================================================
# Recommended preparations for running this script:
# 1. Create a working directory
# 2. Unzip the demo zip-file into the working directory
# 3. Open RStudio from within the subdir "R", by
#    double-clicking "190103_0459AB_PlackettLuce_RelativePaths.Rproj"
# 4. Then only open this script "190103_0500AB_PlackettLuce.R".

# Recommended preparations for the general case:
# 1. Create a working directory
# 2. Inside, create three subdirectories: "data", "output", "R"
# 3. If you wish to make the project portable (relative paths),
#    copy "190103_0459AB_PlackettLuce_RelativePaths.Rproj" into subdir "R"
#    Open RStudio by double-clicking this file.
# 4. Open and adapt script "190103_0500AB_PlackettLuce.R" as desired.
# 5. Create a comma-delimited data file, with the variables for the
#    Plackett-Luce model to be prefixed with "l_" ("l" in honor of "Luce");
#    place it in the subdir "data".
# ========================================================================
# Packages required - install and activate
# ------------------------------------------------------------------------
packages_required <- c("foreign", "PlackettLuce", "tidyverse", "qvcalc")
# Determine new packages to install:
new_packages <- packages_required[!packages_required %in% installed.packages()]
# If any new ones are indeed required, install them:
if(length(new_packages) > 0){install.packages(new_packages)}
# Load / activate the required packages:
lapply(X = packages_required, function(X) library(package = X, character.only = T))
# (lapply = apply to list)
# ========================================================================
# Set paths:
# ------------------------------------------------------------------------
```

```
# path.wd     <- setwd("[User-selected path]")
# Better: Start RStudio with "...\R\190103_0459AB_PlackettLuce_RelativePaths.Rproj"
# This enables relative paths and thus portability of the entire directory.
path.wd     <- getwd()
# working directory: holds this and other project-related R-scripts.
path.base   <- substring(path.wd, 1, nchar(path.wd) - nchar("/R"))
# Relative path to the directory of the project (by clipping off the
# suffix of the path that leads within the projekt folder to the folder
# with the R stuff)
dir.exists(path.base)
# check, if path leads to the correct folder

path.data   <- paste0(path.base, "/data")
# Relative path to input data.
dir.exists(path.data)
# check, if path leads to the correct folder
path.output <- paste0(path.base, "/output")
# Relative path to output data.
dir.exists(path.output)
# check, if path leads to the correct folder
# ============================================================================
# Read data:
# ----------------------------------------------------------------------------
list.files(path.data)
file.exists(paste0(path.data, "/", "181227_2027AB_NPM11_Priorities_PlackettLuce.csv"  ))
data_csv <- read.csv(file = paste0(path.data, "/", "181227_2027AB_NPM11_Priorities_PlackettLuce.csv"),
                     header = T, sep = ",")
# Structure of the data:
str(data_csv)
# Arguments to the Plackett-Luce model have been prefixed "l_".
# All non-numeric data columns are read as factors because
default.stringsAsFactors()
# ============================================================================
# Analyze data:
# ----------------------------------------------------------------------------
# In the needs assessment context, we call the PL-"worth" coefficients
# "Intensities"; this terminology is arbitrary.
data_csv %>%
  dplyr::select(starts_with("l_")) %>%
  PlackettLuce() -> intensities
# The error messages
  # Recoded rankings that are not in dense form
  # Removed rankings with less than 2 items
```

```
# will apppear. They are specific of this dataset and can be ignored.
names(intensities)
#intensities_coef <- coef(intensities)  # Not necessary for the following.
#intensities_coef

# Summary of intensities:
PL_est <- summary(intensities, ref = NULL)
#"ref = NULL" sets the mean of all intensities as the reference value.
# Else the first item would be the reference, with its coefficent constrained to 0.
PL_est # The key output.
# ============================================================================
# Preparations for confidence intervals and exporting output to .csv files
# ----------------------------------------------------------------------------
# Prepares a table for export, stripping the prefix "l_".
# Ensures items will be listed in the sequence of the arguments, not alphabetically.
PL_est$coefficients %>%
  gsub(pattern     = "l_",
       replacement = "",
       x           = row.names(.)) %>%
  factor(x = ., levels = .) -> items
data.frame(items_num = as.integer(items),
           items     = items,
           PL_est$coefficients) -> PL_est_z.values
# PL_est_z.values

# In order to obtain confidence intervals, quasi-variances
# of the coefficients are needed:
qv <- qvcalc(intensities, ref = NULL)
summary(qv)

# Strips prefix "l_" in preparation for export.
# Ensures items will be listed in the sequence of the arguments, not alphabetically.
qv$qvframe %>%
  gsub(pattern     = "l_",
       replacement = "",
       x           = row.names(.)) %>%
  factor(x = ., levels = .) -> items
# Computes additional columns needed for 95%-CIs,
# Exponentiates coefficent and CI bound estimates, in accordance with PL-model.
qv$qvframe %>% # qv table ...
  data.frame(items_num = as.integer(items),
             items     = items,
             .) %>%
```

```
  mutate(quasiSD  = sqrt(quasiVar),
         quasiLCI = estimate - quasiSD * qnorm(0.975, 0, 1),
         quasiUCI = estimate + quasiSD * qnorm(0.975, 0, 1),
         expLCI   = exp(quasiLCI),
         expEst   = exp(estimate),
         expUCI   = exp(quasiUCI)) -> qv_estim_CI
# qv_estim_CI
# ============================================================================
# Export output:
# ----------------------------------------------------------------------------
write.csv(x         = PL_est_z.values,
          file      = paste0(path.output, "/", "PL_coef_table.csv"),
          row.names = F)

write.csv(x         = qv_estim_CI,
          file      = paste0(path.output, "/", "PL_estimates_CI.csv"),
          row.names = F)
# ============================================================================
# Optional end-of-program house-cleaning commands:

# Empties environment completely:
# rm(list = ls())

# Explicit removal of named objects only:
# rm(object_1, .... object_N)
# ============================================================================
# End of script
# ============================================================================
```