CONFIDENCE IN NEEDS ASSESSMENT DATA

The use of confidence ratings in the Syria Multi-Sectoral Needs Assessment (MSNA)

SYRIA

1 APRIL 2015

Aldo Benini
Mohammed Shikh Aiyob
Patrice Chataigner
Benoît Munsch

A note for ACAPS and MapAction

CONFIDENCE IN NEEDS ASSESSMENT DATA

THE USE OF CONFIDENCE RATINGS IN THE SYRIA MULTI-SECTORAL NEEDS ASSESSMENT (MSNA)

1 APRIL 2015

1. SUMMARY

What this is about

Needs assessments strive to collect reliable information. Documenting the reliability adds to transparency and thus is part of the accountability to which humanitarian agencies aspire. Rating and recording the confidence that assessment personnel place in pieces of information is one way of monitoring the reliability of the collected data. The recent Syria Multi-Sectoral Needs Assessment (MSNA) followed this practice in a clearly designed way. Every enumerator who completed his/her data collection in the assigned sub-district was extensively debriefed. Most of the debriefings took place in face-to-face conversations; a minority relied on mobile phone calls with those who could not travel to MSNA coordination offices. The MSNA covered 140 sub-districts and urban sub-divisions and reported findings on 126; Syria has 270 sub-districts.

For each of thirty numeric variables, the enumerators rated the confidence in their estimates. They assigned each non-missing value a confidence rating on a six-point scale. The guidance for determining the confidence level was the same for all thirty variables. Subsequently, enumerators and debriefers together would look at the evidence supporting the estimates. The debriefers would then revise the confidence ratings as they saw fit. All in all, enumerators and debriefers gave 3,666 confidence ratings. These were recorded in the assessment database.

This note is about the larger confidence picture, as evident in the distributions of the confidence scores. It is also about the coherence of the ratings across domains, and about the association of confidence with factors of the conflict environment. It is based entirely on the MSNA's final dataset. Use or non-use of the confidence ratings in team deliberations, analysis and reporting cannot be reconstructed from this basis.

A sidebar in the main part of this note further details the debriefing process. Henceforth, when referring to enumerators and debriefers jointly, we will use the term "evaluators". Readers concerned with the semantics of "reliability" and "confidence" will find the nuances explained in a historic comment, also in the main part.

Variations in confidence

Regarding the distributions of the scores, we find that Level 3 ("medium confidence") is the median for every one of the thirty rating variables. The median, therefore, is not informative, except to say that overall the MSNA team placed a medium degree of confidence in its data. In this situation, a different "good enough" measure of high and low confidence is needed: The sum or, if you will, the proportion in the 126 assessed sub-districts, of the level 1 and 2 cases - those of the highest confidence - serve this function. Using this, we do see some differences in the confidence placed in the thirty variables, but they are not overwhelmingly large. For example, numbers of IDPs in organized camps (where registration may be easier) are better trusted than those in other living arrangements. Sectoral indicators of unmet need are better trusted when they refer to physical structures (houses, schools) than to persons (persons in need). This picture is somewhat obscured by the practice of the evaluators not to rate the confidence in missing values (instead of, appropriately, giving a no-confidence score) and by excluding from the final dataset the records of 14 sub-districts that did not meet minimal standards.

Confidence is one-dimensional

We investigated the dimensionality of the confidence ratings by selecting eight of the 30 rating variables (to minimize redundancy from closely related variables). It turned out that the ratings varied only along one systematic dimension, which accounts for 60 percent of the variability. This means that confidence in the information from all domains - population, IDP locations, sectoral needs - tends to go hand in hand. Because enumerators and sub-districts were in most cases matched one-to-one, it is impossible to say whether the correlations among ratings are due to debriefer bias, enumerator aptitudes, or to objective difficulties in working in the field.

Effects of the conflict environment

However, debriefer bias and enumerator aptitudes are not likely correlated with factors on the ground inside Syria. Thus, any association that we find between confidence and conflict environment should reflect the difficulties of field work (particularly due to insecurity). We tested for the effects of urban vs. rural sub-districts, of recent registrations of persons in need, and of the intensity of recent fighting. We found that

- Urban sub-districts inspire higher confidence in the information that the enumerators brought back, than rural ones do. The effect is not very large, though.
- Similarly, the fact that a sub-district has had a registration of displaced / affected persons in the last three months causes the information to be better trusted. The effect is even much smaller than that of urbanity.
- The level of recent fighting does not have a discernible influence in either direction. This result is likely due to the inaccessibility of the most insecure sub-districts.

Regional differences

Finally, we tested for regional differences. We calculated summary measures for each governorate (the governorates in Syria correspond to provinces in other administrative denominations). The differences in confidence between governorates are less pronounced than those within governorates, i.e. among the sub-districts in a given governorate. Nevertheless there are governorates that overall produced better information, and others

with less trustworthy data. Remarkably, even those with on average higher confidence ratings had some poor-quality returns. Their averages were better because they had a number of very good questionnaires, not because most or all were good. Since no needs assessment should base its findings solely on returns judged of the highest quality, this reinforces the need to deal with information of widely varying quality in ways that are flexible and least exclusionary. It may not be realistic to expect, as a result of better training, supervision and support, to lift the quality of the data across the board. It may be necessary to tolerate a wide range in the quality of the returned assessments, and then to devise finely graded rules for what to do with the weaker ones in analysis and reporting.

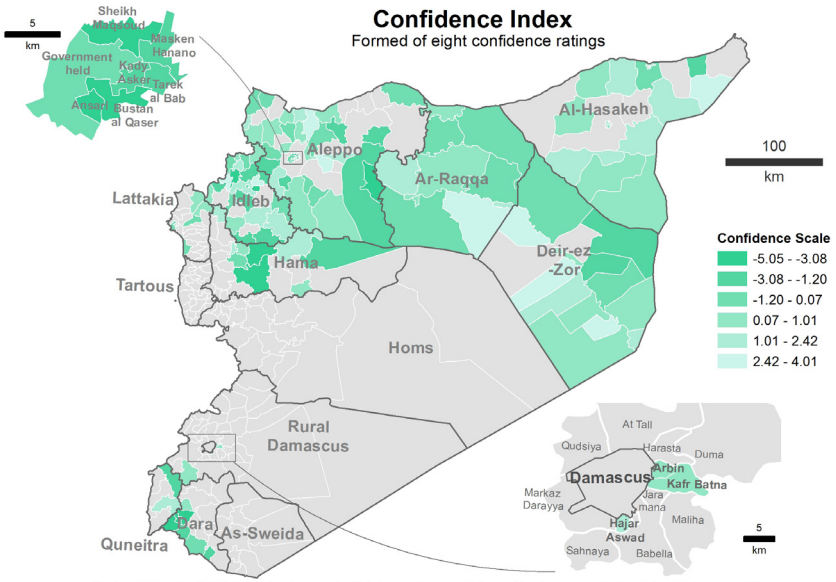


FIGURE 1: Map of the assessed sub-districts, by confidence levels

Note: 126 sub-districts and urban sub-divisions. Lower Values (darker green) stand for higher confidence.

Data Sources: Multi Sector Needs Assessment, October 2014; OCHA COD/FOD registry

The similarity among neighboring units in terms of confidence scores is positive, but relatively low (Moran's I with the three nearest neighbors = 0.15; see appendix.) - The map was produced by Helen Campbell, ACAPS, with input from Jorge Andrés Gálvez.

Limits and recommendations

It is important to show the limits of this study. Confidence ratings are good for discipline, humility and limits of inference. They have downsides, too. They add to the workload. They may create incentives for enumerators to focus on collecting good information on the elements that they know will be rated, at the expense of the quality of unrated elements. Alternatively, some enumerators may overstate the quality of information during debriefings. After debriefings and data entry have been accomplished, the analysts may be able to make very limited use of the ratings, for reasons both of workload / time pressure and consistency of findings in the report.

The recommendations, therefore, are of a modest tactical nature:

- The six-point confidence scale should be continued until new insights emerge. The guidance should be reviewed and perhaps replaced with a simple point system that produces a confidence index, within the range from 1 to 6. The scale should be reversed so that in tables and graphs higher confidence is naturally associated with higher scores. An example is given in the main section on page 9.

- Two data-management recommendations concern the confidence-rating of missing values and the entry of questionnaire returns of low quality into the same assessment database. For the analysis, the working sample of "good enough" records can then be handled flexibly, as appropriate from domain to domain. Some basic information should be presented also about the units (areas, e.g. sub-districts) with low-quality returns; these statistics can be segregated into sidebars or appendices ("Enumerators covered another X districts with an estimated total population of Y thousand people. Due to security problems, the estimates could not be sufficiently verified and are thus not included in the totals presented in the main part of the report. Etc.").

- The type of enumerator training (in-class vs. remote), mode of debriefing (face-to-face vs. by phone), the debriefer's name (or initials) as well as the date of debriefing should be recorded in the assessment database. If enumerators are assigned more than one area, their names (or anonymous identifications) should be recorded as well.

In future, there may be occasions to review the approach to confidence from a strategic perspective. Such occasions may arise on several fronts: from renewed debates on humanitarian accountability, from a stronger penetration of assessment methodologies with statistical concepts of reliability, from social media-supported crowd sourcing that causes a given area (such as a sub-district) to have multiple records in the database, or from other unforeseen evolutions.

2. INTRODUCTION

Reliability is a constant concern in measurement, particularly in data collections of social survey and related types. Conflicts and disasters create turbulent organizational environments; these in turn make for substantial measurement error. This is the type of environment in which humanitarian needs assessment struggle to collect relevant, valid and reliable information. Although statistical concepts of reliability have barely touched this professional community, needs assessment personnel generally collect and process data with an acute concern for quality.

Some needs assessments evaluate their findings less in terms of reliability, and more in terms of the confidence that we can have in the information collected. There may be several reasons for this change in semantics:

- The classic tool for reliability assessments - test/re-test - is practical in humanitarian situations to a very limited degree.
- Rapid needs assessments rely more on expert judgment, provided by secondary data analysts, roving observers and local key informants, than on technical measurement.
- The blurring of lines between humanitarian and security concerns has opened a door for intelligence lingo and thereby for "confidence" talk.

While the concerns are manifest, the practice of reliability and/or confidence ratings in humanitarian needs assessments has not been systematically reviewed. It is not known how they were used in assessments that produced such ratings on a significant segment of the information that they collected. It is even less clear how the variability in ratings of sources, variables or individual data points impinged on findings, and how the assessment reports communicated the uncertainty.

The protracted civil war in Syria has encouraged a habit of self-reflection in a series of needs assessments. Confidence ratings have been introduced, and have evolved with each subsequent assessment (see sidebar below). Most recently, the Syria Multi-Sectoral Needs Assessment (MSNA) (Humanitarian Liaison Group 2014) assessed unmet needs in 140 sub-districts and urban sub-divisions (Syria has a total of 270 sub-districts)[1]. The MSNA design tagged ratings strictly to numeric variables, using the same format for all sectors. The MSNA dataset lets us see at least some aspects of the practice of confidence ratings. This note cannot close the research gap, but it narrows it a tiny bit.

The next sections provide an overview of how confidence ratings are distributed, whether they cluster in one dimension or follow several independent ones. They report tests on whether confidence is associated with factors of the conflict environment. They describe two measures to characterize the aggregate confidence in the data from each assessed sub-district. They compare the regional variation on these measures. We conclude with some recommendations. For the historically interested readers, the appendix recapitulates the evolution of the concept and practice of confidence ratings in needs assessment in Syria in considerable detail.

1 "To account for the diversity of conditions within urban centres, cities were subdivided into smaller units, sectors, to be assessed: Aleppo (7), Deir-ez-Zor (2), Al-Hassakeh (3), Lattakia (2), Damascus (2), Quamishli (2)" (Humanitarian Liason Group, op.cit., page 5). For linguistic simplicity, we will speak simply of "sub-districts".

3. CONFIDENCE RATINGS IN THE MSNA

MSNA analysts extensively debriefed the enumerators who had collected, assembled and were reporting information on the assessed sub-districts in Syria. The debriefings were in part remote (for reasons of security and travel restriction) (46 sub-districts), in part face-to-face (94) (Humanitarian Liaison Group 2014: op.cit., 5). This sidebar details the process and context of the debriefings.

3.1 [SIDEBAR:] DEBRIEFING THE ENUMERATORS

Needs assessments in Syria are viewed with suspicion. Parties to the conflict assume that assessment works are beholden to the other side. The MSNA management, in a bid to strengthen the impartiality of its work, took great care for the quality and credibility of the data as well as of the findings. The systematic debriefing of enumerators was an important element of the quality assurance.

With the debriefings, the MSNA pursued two objectives. The first covered a number of survey quality assurance routines, such as checks for legibility, completeness and consistency. In addition, by marking the places where enumerators had met with key informants on sub-district maps, the debriefers established how completely each sub-district was covered - a rough measure used in determining whether to retain or discard the questionnaire in point.

Second, the debriefers elicited (in Arabic) and noted (in English) qualitative information that would help interpret findings in a properly understood context. For example, an enumerator returned from a relatively small sub-district (current population: 3,800) with an estimate that some 2,500 persons were "in moderate need" for assistance with non-food items (NFI). He described:

> "Regarding NFI some people are using wood to cook, some are borrowing kitchen utensils from their neighbors, buying clothes are luxury now (brothers and sisters are sharing their clothes with each other for example. The community helped as well by providing blankets and anything they can" (Debriefing database, unedited by us).

The observed high degree of mutual assistance may explain the low severity score for the NFI area - the enumerator gave it a "2", meaning a "moderate problem" - as well as the absence of persons deemed in acute need for NFI assistance.

The debriefing database has about 4,500 records, with one text field per record. For an assessment of 126 areas, this is a detailed collection. It is not obvious how much the analysts exploited it for genuine value-added to the findings directly based on the standardized response from the returned questionnaires. Part of the second objective was to collect qualitative information also on domains that the questionnaire could not cover adequately, notably protection. However, it is not clear to what extent the debriefers did so effectively. The assessment report notes (op.cit., page 71) that "in view of these limitations, the protection analysis in this section relies heavily on secondary data, complemented by information from qualitative interviews with enumerators through structured debriefing conversations and focus group discussions with debriefers".

In total 22 persons worked part-time as debriefers. They would sit in pairs with an enumerator, usually for two to two-and-a-half hours. Enumerators who could not come to the coordination offices were debriefed via phone or Skype. After the session, the debriefers would turn in three documents: the cleaned and completed sub-district questionnaire; a table with the standardized response and the confidence ratings for data entry; a write-up of the semi-structured, qualitative interview notes.

Of particular interest is the effect that the debriefings had on the retention vs. the discarding of questionnaires. As we saw, 14 of the 140 returns were rejected. Rejections were triggered by these circumstances:

- The enumerator could not be debriefed.
- The key informant claims were representative of less than 75 percent of the area of the sub-district.
- The debriefing revealed inconsistencies with the known context that the subsequent review by the MSNA coordination team did not resolve.
- The estimates that the enumerator brought back of the IDP and persons-in-need figures deviated by more than 40 percent from those in the so-called Governorates Profile (UNOCHA 2014) AND the Profile estimates were not overridden by those of at least three key informants in the sub-district.

The criteria catalogue indicates that the roles of enumerators, debriefers and, senior to them, the MSNA Coordination Team were well connected. That apart, the requirement to cover 75 percent of the sub-district area - the security challenges being what they were - and the discarding, instead of tagging, of low-quality questionnaires may not have been the best policies.

3.2 GUIDANCE AND PRACTICE

The reported information for the most part was the result of contact with key informants whom the enumerators had talked to while visiting accessible sub-districts. In theory, each trained enumerator was responsible for one sub-district. In practice, 149 enumerators were mobilized for the field data collection to cover 164 geographical divisions; eventually they returned 140 questionnaires.

For thirty of the numeric estimates elicited in the questionnaire, the enumerators and the debriefers (henceforth the "evaluators" when both are included) assigned confidence ratings. The ratings followed uniform guidance across sectoral domains and specific variables; the guidance replaced the previous practice of formulating sector-specific levels of confidence.

This table contains the guidance. Confidence is rated on six levels. The scale is oriented in such a way that "1" designates the highest level of confidence, and "6" the lowest. The confidence ratings were recorded numerically.

Table 1: Confidence rating guidance

Code	Category	Description	
1	Very high confidence level	3 or more different sources of data providing the same exact range of figures. Records available with all the sources and are available for sharing and cross-checking. Records are updated on regular bases. Direct observation matches the data presented and the general opinion of at least 3 people from local population totally matches the data provided. Evidences are available and should explain precise cases (such as photos for all destroyed health centres for instance).	Usable data
2	High confidence level	3 different sources providing a very close range of figures. Records available with at least one of the sources and are available for sharing and cross-checking. Available records are updated on regular bases. Direct observation matches the data presented and the general opinion of at least 3 people from the local population is in line with the data provided. Evidences are available and should explain the general situation (such as photos for all possible shelters of IDPs).	
3	Medium confidence level	1 or 2 key informants provides similar data with limited differences. At least one of them has records but not ready to share necessarily. The records updated on regular bases. Direct observation matches the data presented and the general opinion of at least 3 people from the local population did not show high differences and these people stated a trust with the source of the data. Evidences are not available due to security reasons.	
4	Acceptable confidence level	Only one key informant available on the topic of interest. The key informant has records available but not necessarily ready to share and not being updated on regular bases. Direct observation shows no high differences with the data provided and the opinion of at least 3 people from the local population did not show critically high differences and these people stated a good level of trust with the source. Evidences are not available due to security reasons or other reasons that researchers are supposed to explain during debriefing.	
5	Low confidence level	0-1 key informant available, the key informant has no records. Direct observation shows high differences with the data provided and the opinion of at least 3 people from the local population shows differences or locals stated a low level of trust with the source. Evidences are not available due to security reasons or other reasons that researchers are supposed to explain during debriefing.	Data not usable
6	No confidence	Only one key informant available on the topic of interest. The key informant has no records available. Direct observation shows important differences with the data provided, even if the opinion of at least 3 people from the local population did not show critically high differences and these people stated a good level of trust with the source. Evidences are not available due to security reasons or other reasons that researchers are supposed to explain during debriefing	

Observations rated 5 or 6 were supposed to be discarded in the analysis. This appears to have been done in a combination of single-value exclusion and listwise deletion. Questionnaires with missing or unreliable values (rating > 4) in a number of critical variables that exceeded a set threshold were entirely rejected (listwise deletion). Of the 140 questionnaires returned, 14 were rejected[2].

Individual pieces of data from the retained questionnaires were supposedly excluded from analysis if they were deemed unreliable (single-value exclusion). The extent to which this was actually done is of interest in the specific MSNA context. For example, population information was used on 114 sub-districts deemed sufficiently reliable among the 126 sub-district records retained (page 18 of the report).

It is of lesser concern to this note. We limit it to simply asking whether it was a good thing discarding any information, or whether it would have been more productive to segregate descriptive statistics by levels of confidence. The report might have presented a population summary for the 114 "reliable" sub-districts in the main body. For the twelve included in the report, but not used in the population section, statistics could have been made available in an appendix or sidebar. The fourteen rejects could similarly have been exploited for selective tables clearly set apart. These were options at some point in time; it is now moot to agonize whether they should have been taken.

Our focus is on the distribution of confidence ratings in the 30 rated variables in the 126 retained sub-district records. We test for associations between confidence and potential determinants.

2 The reference to «listwise deletion» is conceptual because we are dealing with options for treating missing and/or low-confidence observations. De facto, the data for 10 of the 14 were never entered.

[Sidebar:] Is the confidence scale guidance optimal?

The guidance is in the column "Description" in the above table. It details the evidentiary requirements for each level. These descriptions are long-winded and indicative, rather than mutually exclusive. Their language is such that they may make it hard on evaluators to quickly determine which level should be assigned to a piece of information.

It appears that the quality of evidence is judged basically on three criteria:

- The diversity of sources (expressed as the number of key informants)
- The degree of agreement among sources (notably in terms of numeric estimates)
- Access to documentary evidence (records, photos).

If this is so, a simple point-based index can be devised, with a range between 1 and 6, as the sum of these responses:

- Number of key informants: 1 → 0 points; 2 → 1 point; > 2→ 2 points
- Disagreement among key informants: None → 2 points; minor → 1 point; major → 0. If only one key informant was available for this piece of information → 1.
- There was relevant documentary evidence: The enumerator actually inspected it → 2 points; it was not accessible, but its existence was credible → 1 point; else 0.

This assumes that there is always at least one key informant as the enumerator's source. It reverses the scale, with the strongest earning 6 points, and with the weakest earning 1. This ordering seems more natural, particularly in charts where highly trusted units will be on the right side of histograms, or on the upper side in two-way graphs when the confidence ratings are plotted on the y-axis.

For more systematic learning, an extended recorded mode may be considered: Instead of simply noting the index value, it may be helpful to provide three small boxes beside each estimate in the questionnaire. In them, the enumerators would report the points for numbers of keys informants, degree of disagreement, and documentary evidence. The index can then be calculated after data entry. This degree of specificity would allow the

assessment team to better locate the sources and limits of confidence. However, this option must be weighed against additional work and training demands.

3.3 DISTRIBUTION OF CONFIDENCE RATINGS

The evaluators rated the confidence in the values of thirty variables. Considering only the 126 retained sub-district questionnaires, they made a total of 3,666 ratings, or 97 percent of the 126 * 30 = 3,780 theoretically possible ratings. This table shows the distributions.

Table 2: Distribution of the raw confidence ratings

Domain / variable	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Ratings
Population							
(Pre-conflict population (2011	19	29	60	18	0	0	126
Persons who fled the sub-district	9	21	72	23	1	0	126
IDPs, total	4	24	77	20	1	0	126
Current population	5	20	73	27	1	0	126
Internally displaced persons							
IDPs in host families	7	13	73	31	1	0	125
IDPs in rented accommodations	6	17	69	33	1	0	126
IDPs in unfinished/damaged buildings	8	17	72	25	1	0	123
IDPs in collective shelters	10	18	75	21	1	0	125
IDPs in organized camps	22	18	62	14	1	0	117
IDPs in self-settled camps	14	20	68	20	1	0	123
(IDPs, total (repeated	7	20	79	19	1	0	126
Food security							
Persons in need of food, acute	9	20	67	27	2	0	125
Persons in need of food, moderate	5	15	77	28	1	0	126
Persons in need of food, all	6	16	72	31	1	0	126
Shelter support							
Damage to houses: none	10	35	59	17	4	1	126
Damage to houses: slight	8	31	60	20	2	0	121
Damage to houses: moderate	8	28	63	20	2	0	121
Damage to houses: heavy	7	27	63	20	2	0	119
Damage to houses: destroyed	9	32	54	19	3	0	117
Health support							
Persons in need of health, acute	12	24	63	19	0	0	118
Persons in need of health, moderate	6	20	69	24	0	0	119
Persons in need of health, all	6	19	70	24	0	0	119
Safe water							
Persons in need of water, acute	12	19	69	23	0	0	123
Persons in need of water, moderate	4	19	71	30	1	0	125
Persons in need of water, all	4	20	71	29	1	0	125
Education support							
Education facilities, damage to: none	14	29	68	11	0	0	122
Education facilities, damage to: slight	12	27	65	15	0	0	119
Education facilities, damage to: moderate	13	25	65	15	0	0	118
Education facilities, damage to: heavy	12	29	61	14	0	0	116
Education facilities, damage to: destroyed	15	24	60	13	0	0	112
Total ratings for given confidence levels	**283**	**676**		**650**	**29**	**1**	**3,666**

A number of patterns can be fleshed out:

- The majority of the rated information earned a "medium confidence" rating, which the guidance defines by a modicum of consistency among local sources and/or the quality of the records that the enumerators saw.
- Level 5 and 6 ratings were rarely given. This must be so because questionnaires given many such ratings were entirely disqualified (the 14 listwise deletions). Had the ratings of the rejected questionnaires been recorded, we might see a fairly symmetrical distribution around the medium confidence category.
- Level 3 is the median category of every one of the 30 distributions, and thus the row medians in this table do not discriminate. For a quick first measure of the overall confidence in a variable, the sum of the level-1 and -2 frequencies is informative. Using this, we find that confidence varies among the variables within some domains. Not surprisingly, the pre-conflict population figures enjoy higher confidence than the estimates of current demographic variables. Numbers of IDPs in organized camps

(where registration may be easier) are better trusted than those in other living arrangements. Persons in acute need are estimated with greater confidence than those in moderate need[3].

- There are differences in confidence between domains as well. Sectoral indicators of unmet need are better trusted when they refer to physical structures (houses, schools) than to persons (persons in need).
- Overall, the numbers of missing ratings are small, but their variation across the thirty variables matters. The evaluators did not rate confidence in instances where the underlying variable itself had a missing value. It can be shown that variables enjoying relatively high confidence (on this quick first measure) tend to be those with more missing ratings. In our example above, it is precisely the variable "IDPs in organized camps" that had the most missing.

This confounds any conclusions about which variables earned greater confidence overall (since, in the logic of the quality of evidence, we would assign a missing value a confidence level of 6 - or worse!). One can speculate that two things happened: 1. Zeros in some underlying variables (e.g., the sub-district had no IDPs in organized camps) were entered as blanks; 2. Some evaluators preferred setting dubious values to missing, rather than assigning them poor confidence ratings. For our later analyses, we will work around this difficulty. Depending on the type of analysis, we will recode all missing ratings as "Level 7", or recode 5, 6 as well as missing all as 4.

3.4 DIMENSIONS OF CONFIDENCE

Is the confidence that the evaluators placed in a given value independent of the confidence given to the values in other variables? Or are they closely associated with each other, to the effect that information about some sub-districts enjoyed consistently higher trust than that about others?

We pursue two approaches in order to elucidate this question.

- One may assume that confidence is not systematically tied to domains, but is determined chiefly by perceptions of how difficult it was for enumerators to work in given sub-districts. If so, the up to 30 confidence ratings that an evaluator noted in the sub-district questionnaire expressed the same undifferentiated confidence. If he considered the sub-district difficult, he had low trust in the estimates across the board. If working conditions were less restrictive, confidence would have been universally higher. Debriefer attitudes may have had a similar effect. If the debriefer had a high opinion of the enumerator's ability, he was likely to accept favorable initial ratings, or even improve on unfavorable ones. A low opinion would have had the opposite effect. In all those scenarios, all of the 30 confidence ratings can be considered Likert items, and a Likert-scale model (Wikipedia 2011), in which the ratings are treated as interval-level data, seems appropriate. Practically, we set the missing ratings to 7, and then calculate for every sub-district the mean of the 30 ratings.
- The opposite view holds that evaluators feel strongly that the differences in the confidence in the information from different domains are genuine, based on detailed reasoning. The question then becomes one of how many distinct dimensions their confidence takes when viewing the ratings on all 126 sub-districts together. Since the ratings, in this view, remain ordinal, our analytic options are limited. We also have to avoid clustering among the variables within given domains (which might produce spurious statistical factors). The way around is to select one variable from each domain, with the exception of demography, where it seems justified to consider the confidence placed in the estimates of the pre-conflict as well as the current population. We then subject the eight selected variables to a kind of correlation analysis that is appropriate for ordinal data. Technically interested reader may consult the appendix.

The histogram below presents the result of the first approach. It places each sub-district by its mean confidence rating within a range of 0.5. The mean of means is 2.98, almost exactly the numeric value of the median-confidence category. The observed range runs from 1 to 4.7. The tail on the right-hand side is thinner than it would be had the ratings been recorded from the 14 rejected questionnaires. With this proviso, we may believe that the evaluators, in assigning confidence ratings, were following a mental model of "some good, many ok, a few poor". If we apply cuts at the mid-points 2.5 and 4.5, we find 5 + 6 + 17 = 28 good returns, 27 + 46 + 15 + 8 = 96 that are ok, and 2 + the 14 excluded = 16 of poor quality. From a "good-enough" philosophical angle, one can work with such a portfolio.

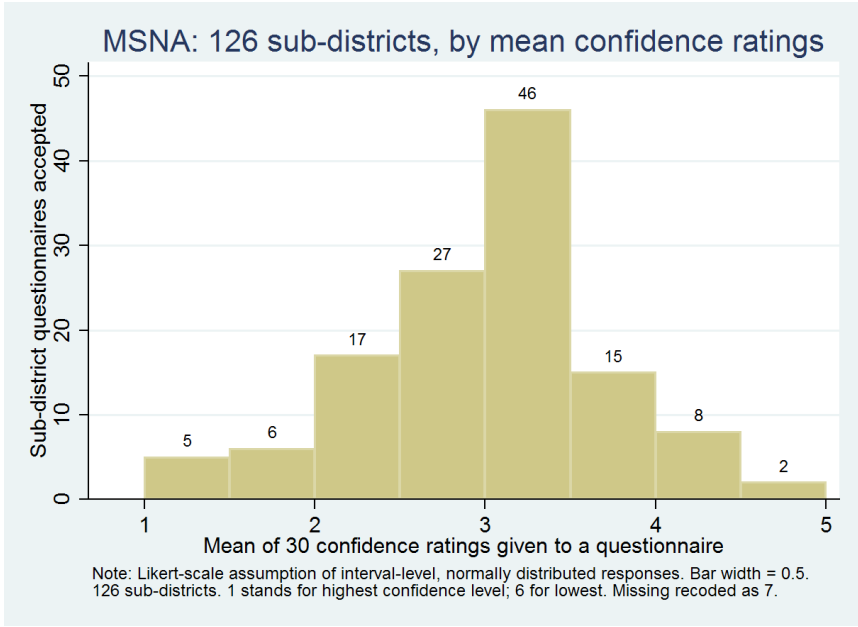


FIGURE 2: Sub-districts, by confidence in the assessment information

3 The distinction between persons in acute need and those in moderate need is the topic of a separate note.

The second type of analysis has a clear result. There is only one systematic component in the confidence ratings. High confidence in the values of some variables encourages high confidence in the others, and vice versa. Because enumerators and sub-districts were matched one-to-one, it is impossible to say whether that reflects general trust in the personality of the enumerator, across all domains. It may just as well reflect that working in some sub-districts is easier than in others, in whatever domain. The dominant dimension (or, as we say technically, "principal component") accounts for 60 percent of the ratings' variability. In other words, 40 percent are unique to the particular enumerator/sub-district, or result from faulty ratings. For illustration, here is an example of how two of the confidence variables are associated. The ratings concern the confidence in the number of IDPs in organized camps and in the estimated proportion of houses that have suffered no damage:

Table 3: Example of two confidence variables, cross-tabulated

Confidence, IDPs in organized camps	Confidence, damaged houses: none were damaged						Total
	1	2	3	4	5	6	
1	9	5	6	1	1	0	22
2	1	10	6	1	0	0	18
3	0	13	34	11	3	1	62
4	0	2	9	3	0	0	14
5	0	0	1	0	0	0	1
Total	10	30	56	16	4	1	117

Had the confidence ratings of the 14 rejected questionnaires been recorded, we would probably see cases in the lower right corner - where confidence is low in both variables -, and all the more so if missing values were re-coded as extremely low confidence.

Since the second model returned only one systematic dimension, we are not surprised to find that the two measures - the sub-district's mean over the 30 ratings and the score on the principal component are highly correlated (+0.88). Because of this high correlation, we would learn little from visualizing the score distribution beyond what the histogram above tells us. The "uni-dimensionality" of confidence ratings leaves us at a loss as to what caused it:

- Debriefer bias (the debriefer liked some enumerators, but discounted the quality of the work of others)
- Differences in enumerator aptitudes (some delivered better work across domains, others much less so)
- Difficulties while in the assigned sub-districts (for security reasons, collecting information from key informants was easier in some sub-districts than in others).

Debriefer bias and enumerator aptitudes are not likely correlated with factors on the ground inside Syria (with the exception that some enumerators were debriefed over the phone because they could not travel. Remote debriefings may have produced different confidence ratings from those face-to-face.). Therefore, if in the next section we find systematic differences in confidence dependent on conflict-related factors, we may attribute them, in large part, to more or less challenging enumerator work conditions.

3.5 CO-VARIATES OF CONFIDENCE

We have no substantive theory of what factors in the conflict environment boost or inhibit confidence in the information that the enumerators brought back. At most, we can formulate some ad-hoc hypotheses, on a common-sense basis:

- Urban sub-districts inspire greater confidence in the information than rural ones do. It may be more difficult for enumerators to travel to the four corners of a far-flung rural sub-district, and therefore the key informants that they meet in one or two locations are not seen as representative of the entire community. If an enumerator in fact manages to criss-cross the area extensively, he may find that the key informants describe situations in very localized ways. If they extrapolate their estimates to the sub-district, the enumerator will likely find large discrepancies. Urban areas, due to shorter travel distances, may pose lesser challenges in this regard; the key informants may have fuller knowledge.

- The ability of sub-district-based organizations to register persons in need may indicate a higher level of administrative penetration within the affected population, a condition that helps enumerators find, and work with, knowledgeable key informants. The MSNA asked about recent registrations in a general way: "Have the displaced/crisis-affected people been registered in this sub-district in the last three months?" Apart from the specific statistics that the registration efforts produced, the registration may have been the work of persons who had information and insight beyond the covered items.

- The level of recent fighting in the sub-district would have the opposite effect. Fighting disrupts the lives of the affected persons. The disruptions weaken the knowledge that key informants have been forming of the situation, especially in terms of numeric estimates of affected groups. Persons organizing local relief, or tending to the wounded, may revise subjective estimates continuously. But under fire, numeric estimates will be less accurate and precise than those arrived at in quiet periods.

We test the three bullet-pointed hypotheses in this way:

We tabulate each of the 30 confidence rating variables by the hypothetical factor. For each table we calculate a measure of association between the ratings and the factor values (Goodman-Kruskal's gamma; see appendix for technical details). We treat the measure as a random variable of which we have 30 observations, within a range of -1 to + 1, meaning: from perfect negative to perfect positive association. If the factor has a consistent effect on the confidence, then the mean of the association measures should be different from zero to a statistically significant degree.

We find that:

- Urban sub-districts inspire higher confidence in the information that the enumerators brought back, than rural ones do. The effect is not very large (mean gamma = +0.18[4]).

- Similarly, the fact that a sub-district has had a registration of displaced / affected persons in the last three months causes the information to be better trusted (this does not apply to IDP-specific variables only, but to the ensemble of 30 confidence-rated variables). The effect is even much smaller (gamma = -0.05[5]).

- The level of recent fighting does not have a discernible influence in either direction.

The results lend some credence to two of the three hypotheses. But the significance is purely statistical. The effects are too small to justify policy changes.

3.6 REGIONAL DIFFERENCES

Finally, we anticipate regional differences in the confidence with which the sub-district information was received. It is not possible to hypothesize the direction of these regional effects; there is no guiding theory. The comparison, therefore, is purely exploratory. We compare the distribution of the confidence measures among the governorates, and also the variation within governorates. The next table sorts governorates by the mean of the 30 ratings-based measures. The subsequent composite graph shows the position of each sub-district within a given governorate on both measures.

With the exception of Dar'a and Ar-Raqqa governorates, the averages differ by little. Dar'a is the only one that inspired distinctly higher confidence in the data. But one needs to know that the debriefing of all the enumerators in Dar'a was done over mobile phones. This may have encouraged shorter sessions than was typical of face-to-face debriefings, with a more lenient attitude on the part of the debriefers, giving the benefit of the doubt to the enumerators. Ar-Raqqa's mean was pulled up on the scores (= pulled down in confidence) by three weak sub-district returns, but its median is still 3 (Dar'a's is 2). All in all, between-governorate differences are minor.

Table 4: Sub-district confidence ratings, mean by Governorate

Governorate	Sub-districts assessed	Confidence measures	
		Mean of 30 ratings	Principal component, based on 8 ratings
Higher confidence:			
Dar'a	7	2.15	-2.05
Idleb	23	2.82	-0.25
Hama	9	2.85	-0.49
Lattakia	5	2.87	-0.17
Aleppo	35	2.91	-0.50
Lower confidence:			
Al-Hasakeh	16	3.22	0.84
Quneitra	2	3.22	-0.25

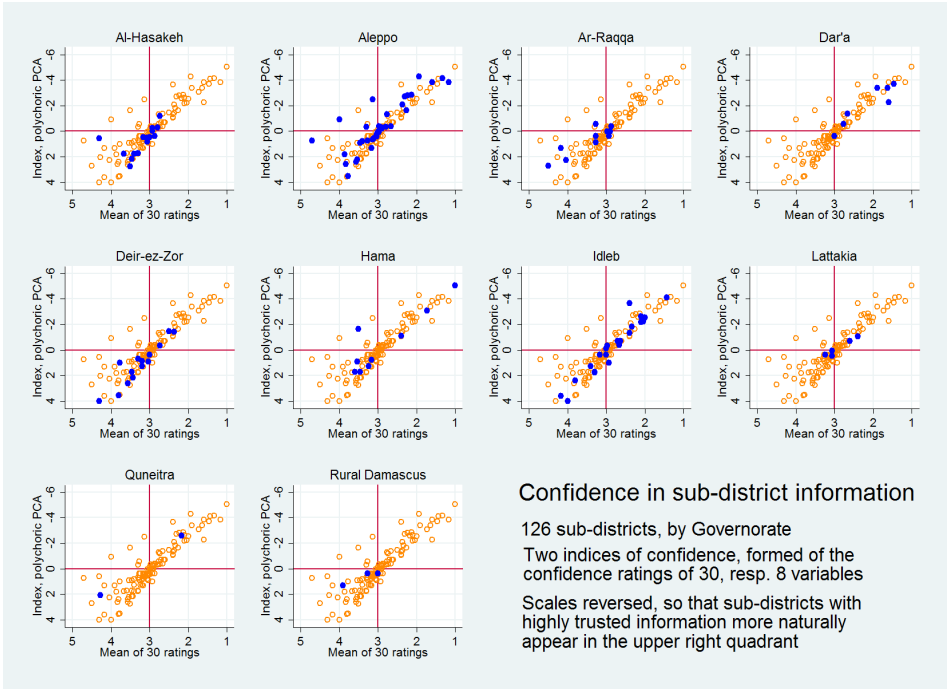
4 The rural/urban variable is coded 1 for «rural», 0 for «urban». Higher values on the confidence scale mean less confidence (1 is best, 6 is worst). The positive association measure says: Rural sub-district tend to have (somewhat) higher values on the confidence scale, i.e. are less trusted. This is equivalent to the (easier) reformulation in terms of «urban .. higher confidence» in the text.

5 Registration is coded as 0 «No or not yet», 1 «Under way», 2 «Completed».

Deir-ez-Zor	15	3.23	1.05
Rural Damascus	4	3.29	0.61
Ar-Raqqa	10	3.42	0.70
Total	**126**	**2.98**	**-0.04**
Note: Sorted on the measure based on 30 ratings			

The chart highlights the variation within governorates. Each panel marks the positions of the assessed sub-districts with blue dots, on the backdrop of the sub-districts of all other governorates represented by orange circles. The upper right quadrant is home to sub-districts with data enjoying above-average confidence; the lower left is for the below-average. The other two quadrants hold a small number of sub-districts that are above-average on one measure, and below on the other.

FIGURE 3: Within-governorate variation of confidence measures



If we leave aside the governorates with few assessed sub-districts (< 9), we find an interesting relationship between the mean and the variance of the sub-district ratings:

Aleppo, Hama and Idleb not only have lower means (i.e., on average enjoy higher confidence) than Al-Hasakeh, Ar-Raqqa and Deir-ez-Zor; their blue dots are also farther spread out. In other words, they achieved higher confidence in the quality of their data not because there were no sub-districts delivering low quality, but because they also had a number of very good ones. This finding may be of some policy interest: It may not be realistic to expect, as a result of better training, supervision and support, to lift the quality of the data across the board. It may be necessary to tolerate a wide range in the quality of the returned assessments, and then to devise finely graded rules for what to do with the weaker ones in analysis and reporting.

Overall, the findings about regional variations are not dramatic. The differences within given areas seem more important than those between areas. The outlier case of Dar'a governorate suggests that debriefing by phone may inspire higher confidence in the data, but this is far from certain because the medium of debriefing the enumerators who worked elsewhere was not recorded in the database.

4. CONCLUSIONS AND RECOMMENDATIONS

4.1 CONCLUSIONS

The MSNA went to great lengths not only to ensure good reliability, but also to document it. It couched this quest in the vocabulary of confidence, rather than of reliability. This is justified because needs assessments can hardly ever afford multiple independent measurements of the same items. The practical device for documenting the confidence in the collected information was the extensive debriefing that core team members administered to returning enumerators. The enumerations rated their confidence in each of the returned data points in as many as thirty variables. These were all numeric and were evaluated using the same six-level confidence scale. The debriefers accepted or modified the enumerators' ratings on the strength of the evidence.

The confidence ratings were recorded as an integral part of the database, which makes them transparent and open to analysis. This is a great strength of the MSNA methodology, in contrast to needs assessments that do not differentiate levels of confidence or other quality measures. Regrettably, however, the MSNA excluded the data on 14 sub-districts on account of their perceived poor overall quality. This creates a selection effect, which hampers the analysis and ultimately the conclusions and recommendations that this note can make on the basis of the final dataset

Although there is considerable variation in the confidence ratings, overall the differences are modest. This is true in several regards. The differences follow expectations for the most part, by type of variable, by clustering among the ratings of different variables, and by association with factors of the conflict environment. The potential for large differences is dulled by the simple fact that in every one of the thirty rated variables, level 3 is the median confidence level.

If the median is the same for all, a useful heuristic is to take the sum of cases (sub-districts) in the highest two levels of confidence, coded as level 1 and 2. Here some interesting differences appear, particularly within families of related variables. For example, estimates of IDPs enjoy higher confidence when the IDPs are located in places clearly demarcated from the host population. Conversely, estimates of IDPs in host families and in rented accommodations seem less trustworthy. But, again, the differences are relatively minor. They are also confounded by the practice of not giving a confidence rating for missing observations, instead of assigning the code expressing the least trust.

One might expect that enumerators would have produced trustworthy estimates in some domains while failing to gather strong evidence in others. In this view, which domains offered the stronger evidence depended on luck and local circumstances. But this is so to a minor degree only. Overall, the confidence ratings across domains are strongly correlated, such that there is only one systematic dimension. Sub-districts releasing information of good quality in one domain tended to do so in others, and vice versa. Because in most cases enumerators and sub-districts were matched one-to-one, it is impossible to say whether that reflects debriefer bias, enumerator aptitude or objective difficulties in the field, notably of the security kind.

Two of the three outside factors - rural/urban, recent registrations, but not recent levels of fighting - tested positive for associations with the majority of the thirty confidence variables. Urban sub-districts inspire more confidence, and so do sub-districts that recently saw some kind of registration of persons in need. But, again and again, we should stress that these effects are modest. The absence of a clear effect of recent fighting is likely due to restricted access to the most insecure areas.

A finding regarding regional differences merits repetition. The differences in confidence between governorates are less pronounced than those within governorates, i.e. among the sub-districts in a given governorate. Nevertheless there are governorates that tended to produced better information, and others with less trustworthy data. Remarkably, those with on average higher confidence ratings too had some poor-quality returns. They were better because they had a number of very good questionnaires, not because most or all were good. Since no needs assessment should base its findings solely on returns judged of the highest quality, this reinforces the need to deal with information of widely varying quality in ways that are flexible and least exclusionary.

There are a number of practical questions that this note cannot address. Confidence ratings add to the workload. They may create incentives for enumerators to focus on collecting good information on the elements that they know will be rated, at the expense of the quality of unrated elements. Alternatively, some enumerators may overstate the quality of information during debriefings. After debriefings and data entry have been accomplished, the analysts may be able to make very limited use of the ratings, for reasons both of workload and consistency of findings in the report.

4.2 RECOMMENDATIONS

The MSNA operated with six confidence levels and with a uniform guidance applicable to all variables that the evaluators rated. There is not enough evidence to recommend changing the number of levels making up the confidence scale. The fact that levels 5 and 6 were rarely used is due to the censoring of the 14 rejected questionnaires - this is a different problem (see below). Until new insights emerge, the six-level scale should be continued.

However, the guidance for the administration of the six-level scale should be reviewed. A simple point-based system, outlined on page 9, may be an option for greater clarity and quicker decisions.

More importantly, two changes are recommended in the ways the assessment data are processed:

1. Missing values in confidence-rated variables need to trigger ratings. This could be the same level as the least trusted information (level 6 in the MSNA set-up), or even a worse level reserved to missing (in this set-up, missings would be confidence-coded as 7). This goes hand in hand with the discipline of distinguishing between zero and missing. For example, if a sub-district has no organized IDP camps, then the IDP population in organized camps is zero. It is zero, not missing. If we are absolutely positive that there are no such camps, then the zero population data point enjoys the highest confidence (level 1 in the MSNA). However, if there is a reasonable assumption that such camps exist, but none of the sources offers a count or estimate of the IDPs living there, we have a situation of missing, and need to assign the confidence code for missing.

2. Qualitatively weak questionnaires should be entered into the database, together with the confidence ratings. In the analysis, the way of segregating the acceptable from the poor records is to create tagging variables. Tagging variables take the value of 1 if the record is to be included in the working sample, and 0 if not. The logic of tags has been described in an earlier ACAPS note "A template for managing data in needs assessments" (Benini 2012). In Excel-supported analyses, tags are used as filters in Pivot tables. Tagging variables are flexible; they can be filled manually for each record, or they can be computed as any function of a set of confidence rating variables. There can be as many tagging variables as makes sense analytically, perhaps a different one for each sector, as the analyst wishes. The great advantage is that the records (sub-districts in the MSNA) that are excluded from the main analyses are kept available. Since their information may have value in some contexts, statistics on them are easy to produce and can (and should) be presented in sidebars or in appendices.

The mode of debriefing - face-to-face, by phone - may impact the confidence in the information. There may also be collective learning effects over time that change the way of determining confidence levels between the earlier and the later returns. The mode, the debriefer's name (or initials) as well as the date of debriefing should be recorded in the assessment database. If enumerators are assigned more than one area, their names (or an anonymous identification) should be recorded as well.

These recommendations are tactical. There may be occasions to review the approach to confidence from a strategic perspective. Such occasions may arise on several fronts: from renewed debates on humanitarian accountability, from a stronger penetration of assessment methodologies with statistical concepts of reliability, or from social media-supported crowd sourcing that causes a given area (such as a sub-district) to have multiple records in the database, or from other unforeseen evolutions.

1. APPENDIX

Tables and graphs in the appendix are not captioned.

1.1 THE EVOLUTION OF CONFIDENCE SCALES IN NEEDS ASSESSMENTS IN SYRIA

Historically, the quality of information collected during field assessments in emergencies was assessed on two criteria: the reliability of the source as well as the credibility of the content. We find this concept adopted as early as the year 2000, explicitly in the UNDAC Field Handbook. Its latest (sixth) edition (UNOCHA 2013: 5) formalized six levels on both criteria:

Reliability of source	Credibility of information
A. Completely reliable	1. Confirmed by other sources
B. Usually reliable	2. Probably true
C. Fairly reliable	3. Possibly true
D. Not usually reliable	4. Doubtful
E. Unreliable	5. Improbable
F. Reliability cannot be judged	6. Truth cannot be judged

In this scheme, each piece of information receives an alphanumeric rating reflecting the enumerator's level of confidence. For example, an element obtained from a source regarded as "usually reliable" and deemed "probably true" will trigger a B2 rating.

Varieties of reliability and credibility ratings have been adopted by other humanitarian actors such as in the WFP's Integrated Phase Classification System and by UNICEF. Reliability and credibility criteria are used also in other fields that struggle with incomplete information and decision-making under time pressure, notably the military and the intelligence services. The admiralty grading system in the UK government and the source reliability rating matrix in the Joint Intelligence Manual of the Canadian Government are examples in point (Hibbs-Pherson and Pherson 2012; Wittek and Zwitter 2014).

Efforts to systematize quality control in needs assessments in Syria go back to 2013. The practice of confidence ratings has evolved over five assessments so far. A remarkable stability among key members of the coordination teams over time has encouraged significant organizational learning. This table details the scope of confidence ratings and the scales used in successive assessments.

Assessment	What was rated?	Scales	Thresholds for exclusion
J RANS 1 (ACU 2013)	Population figures	Rating: 1=reliable, 2=fairly reliable, 3= unreliable	Information with a level 3 confidence rating was discarded from final analysis
J RANS 1.5 (Aleppo) (AWG 2013a)	Population figures	Rating: 1=reliable, 2=fairly reliable, 3= unreliable	
J RANS II (AWG 2013b)	Population figures Damage levels Humanitarian access All the data relating to an entire sector	**Evidence Rating**: 1. Strong evidence – verified by enumerator or very credible sources, triangulation between different sources confirms **same**, observation confirms findings; 2. Good evidence – triangulation between different sources confirms **similar**, credible sources; 3. Triangulation not possible or sources not credible or triangulation reveals significant differences, information not confirmed with evidence, no observation.	
SINA (AWG 2013c)	Population figures Sector-wise persons-in-need figures	**Reliability Rating**: 1. Strong reliability – verified by enumerator or very credible sources, triangulation between different sources confirms **same**, observation confirms findings; 2. Good reliability – triangulation between different sources confirms **similar**, credible sources; 3. Triangulation not possible or sources not credible or triangulation reveals significant differences, information not confirmed with Reliability, no observation.	
MSNA	Population figures Sector-wise persons-in-need figures	**Confidence level:** 1.Very high confidence level, 2.High confidence level, 3.Medium confidence level, 4.Acceptable confidence level, 5.Low confidence level, 6.No confidence	Information with a level 5 and 6 rating was discarded from the final report

The multiplicity of multi sectoral assessments in Syria offered an opportunity for trial and error in 2013 and 2014.

1. The first J-RANS (January and February 2013) used a simplified three-point scale. At this time, humanitarian data collections within Syria were relatively recent and few; thus the enumerators had little to go on for triangulation (local committees were not yet active in many places; registration were just being started). As a result, a large number of records were discarded. In fact, in J-RANS 1, more than two million affected persons were excluded from the humanitarian profile because the information gathered on them appeared unreliable.

2. J-RANS II refined the criteria for rating evidence. The ratings were expanded to more sections of the questionnaire. Of special note, entire sections, each covering a sector, were made the objects of ratings (meaning that one rating was determined for the entire sector information). This scale opened the way to more systematic debriefings. However, the use of one single evidence rating for an entire set of questions had unintended consequences: a rating of "3" would cause the entire section to be discarded. A lot of information could have been saved for the analysis had the confidence ratings been aimed at specific questions within the sector sections.

3. SINA learned the lessons from the J-RANS. It restricted reliability ratings to critical variables, notably the population figures and the estimates of people in need at sector level. By this time registrations at sub-district level were more common; as a consequence, the proportion of discarded data was much lower than in the previous J-RANS. However, the accuracy of the estimates was severely challenged during the validation process. In particular, critics questioned the claimed numbers of key informants interviewed.

4. Finally, nearly one year after the SINA, the MSNA opted for confidence levels. The term "confidence" was chosen in line with publicized intelligence analysis techniques.

The intelligence community is wont to evaluate several aspects of an information ensemble - sources, content and the conclusions that the information appears to support.

Level of aggregation	Reliability of the source	Credibility of the information	Confidence in the conclusions
One evidence/data point			
Several pieces of evidence ((related to the same topic		Requires several pieces of evidence for triangulation	
Conclusions drawn from multiple pieces			Requires aggregation and interpretation

These distinctions guided also the confidence rating process that the MSNA adopted. Its enumerators had to aggregate information from multiple sources and places. The best estimates that they would report were the result of deliberate interpretation. The multiplicity of sources as well as the dependence on context and local knowledge suggested that the term "confidence" was better suited to describe the process that led to the final estimates.

In practice, the confidence scores were determined by deliberations that took into account several factors:

1. The strength of the evidence, based on the reliability of sources and credibility of the information (from uncorroborated to well corroborated information)

2. The number and importance of key assumptions or adjustments used to fill information gaps (from many to minimal assumptions or adjustments).

3. The strength of the underlying logic, measured in part by the aggregation methods (from weak to strong logical aggregation or inferences), plausibility and consistency across the questionnaire (i.e. there could not possibly be more people in need than people affected)

4. The agreement between the enumerator and the debriefer.

The MSNA adopted a six-point scale, compared to the three points common to the preceding assessments. It did so for several reasons:

- Evaluating the overall quality of evidence, with a view to forming realistic expectations in future assessments in Syria. In other words: what are reasonable data quality standards in contexts like Syria?

- A refinement of the number and combination of criteria for judging the reliability and credibility of the information (both for the enumerators and the debriefers), based on the following parameters:

Reliability of the source	Credibility of the information
• Professional competence of the source • Motives for bias • Reputation, track record of accuracy and level of trust	• Number of key informants interviewed • Records and documentation availability • Frequency of updates, and recency of the latest update • Match with findings from direct observation • The degree to which results can be confirmed or corroborated by other type of evidence (photos, videos, testimony, secondary data, etc.) • Level of agreement or plausibility with other people's opinions

As a result of the new confidence scale and of the greater refinement of the evaluation criteria, relatively few numeric data points were disqualified from the final MSNA analysis (approx. 30 out of a total of 3,666 data points).

1.2 STATISTICAL ANALYSES

Polychoric Principal Components

The confidence ratings are not independent the ones of the others. High confidence in the values of some variables in a given sub-district tends to go hand in hand with similarly high confidence in those of others. For the reasons mentioned in the main text, we selected eight variables from whose correlation pattern we sought to establish on how many major underlying dimensions they clustered. For ordinal variables, as these are (outside the Likert scale interpretation), polychoric principal component analysis (Kolenikov and Angeles 2004) is the appropriate procedure. The scores of the first component are the values of the confidence index displayed in the map in the summary.

Descriptive statistics

variable name	storage type	display format	value label	variable label
A1_a_4	byte	%10.0g		Confidence, pre-conflict population (2011)
A1_d_4	byte	%10.0g		Confidence, current population
A2_e_4	byte	%10.0g		Confidence, displaced in organized camps
C2_a_2	byte	%10.0g		Confidence, PiN food, acute
D4_a_2	byte	%10.0g		Confidence, damage to houses: none
E2_a_2	byte	%10.0g		Confidence, PiN health, acute
F2_a_2	byte	%10.0g		Confidence, PiN water, acute
G8_a_2	byte	%10.0g		Confidence, educ facil damage, none

There were few ratings of levels 5 and 6 and relatively few missings (Confidence in the total number of IDPs had the highest number, 9 out of 126). For stable estimation, we re-coded 5, 6 and missing as 4. The recoded rankings were distributed thus:

Variable	Values 1	2	3	4	Total
conf1to4_A1_a_4	19	29	60	18	126
conf1to4_A1_d_4	5	20	73	28	126
conf1to4_A2_e_4	22	18	62	24	126
conf1to4_C2_a_2	9	20	67	30	126
conf1to4_D4_a_2	10	35	59	22	126
conf1to4_E2_a_2	12	24	63	27	126
conf1to4_F2_a_2	12	19	69	26	126
conf1to4_G8_a_2	14	29	68	15	126
Total	103	194	521	190	1,008

The resulting PCA estimate returned a single component with an eigenvalue > 1. It accounts for 60 percent of the variability.

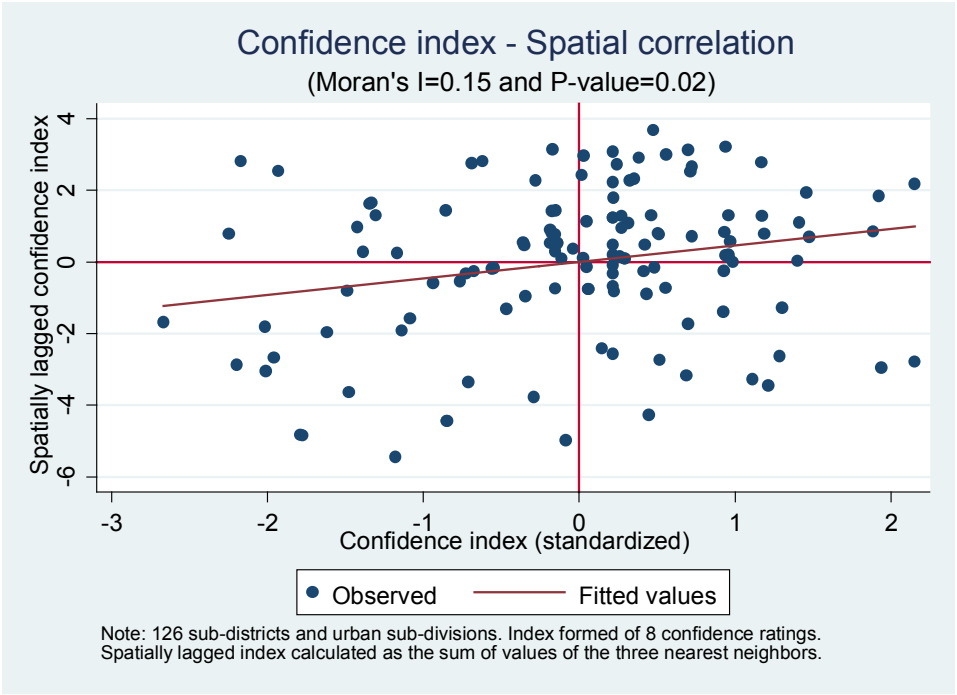
Principal component analysis

k	Eigenvalues	Proportion explained	Cum. explained
1	4.813307	0.601663	0.601663
2	0.891493	0.111437	0.713100
3	0.674364	0.084295	0.797395
4	0.484148	0.060518	0.857914
5	0.405519	0.050690	0.908604
6	0.278342	0.034793	0.943396
7	0.232653	0.029082	0.972478
8	0.220175	0.027522	1.000000

Although there is only one systematic component, the 60 percent variability is relatively low, too low to suggest that the Likert scale model is fully appropriate. The proportion might have been higher if the ratings of the rejected questionnaires had been recorded.

Spatial correlation of the confidence score

We investigated the dependency of the score on the scores of neighboring sub-districts. Since the assessed sub-districts do not form a contiguous region, we relied on distance between centroids rather than on adjacency for the selection of relevant neighbors. Arbitrarily, we defined the three sub-districts with the shortest distances between their respective centroids and that of the sub-district in focus as its relevant neighbors. This is not an ideal situation because there are small clusters and isolates for which the all or some of the three nearest neighbors have no objective relevance for this question.



Note: 126 sub-districts and urban sub-divisions. Index formed of 8 confidence ratings. Spatially lagged index calculated as the sum of values of the three nearest neighbors.

We used the Stata procedures "spwmatrix" to create the nearest-neighbor matrix and "splagvar" to compute Moran's I as well as the corresponding scatterplot (Jeanty 2010a, 2010b). About Moran's I and other measures of spatial association, see Anselin (1995) and Wikipedia (2015).

Tests for effects of conflict factors on confidence levels

As noted, the confidence scale is coded 1 for the highest level, and 6 for the lowest. For the purpose of these tests, we re-coded missing as 7 (no confidence whatsoever).

The measure of association chosen between confidence scale and conflict factors is Goodman and Kruskal's gamma (Wikipedia 2014). Gamma does not adjust for ties; but ties are expected since the number of levels is minimal (some variables such as rural/urban are dichotomous). Kendall's tau-b, adjusting for ties, is generally closer to zero than gamma, but for the question whether the mean measure over the 30 confidence ratings is different from zero this is unlikely to change results (we did not replicate tests using tau-b).

Rural /urban

```
variable name   type    format     label      variable label
------------------------------------------------------------------------------
isRural         byte    %8.0g      isRural    Rural vs. urban sub-district

isRural:
           0 Urban
           1 Rural


  Rural vs. |
      urban |
sub-dist.   |
            |      Freq.     Percent        Cum.
------------+-----------------------------------
      Urban |         48       38.10       38.10
      Rural |         78       61.90      100.00
------------+-----------------------------------
      Total |        126      100.00

variable name   type    format     label      variable label
------------------------------------------------------------------------------
GammaMiss_isR~l float   %9.0g                 Gamma Rural sub-district vs.
                                              confid. ratings (missing
                                              recoded 7)

. summ GammaMiss_isRural, detail

                    GammaMiss_isRural
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.0657222      -.0657222
 5%    -.0653222      -.0653222
10%     .0124082       -.010813       Obs                  30
25%     .1126878       .0356295       Sum of Wgt.          30

50%     .1799458                      Mean           .1849354
                       Largest        Std. Dev.      .1293242
75%     .2602947       .3674093
90%     .3809269       .3944444       Variance       .0167248
95%     .4164134       .4164134       Skewness        .055456
99%     .4278238       .4278238       Kurtosis        2.65057

. ttest GammaMiss_isRural == 0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
GammaM~l |      30    .1849354    .0236113    .1293242    .1366449    .2332259
------------------------------------------------------------------------------
    mean = mean(GammaMiss_isRural)                               t =   7.8325
Ho: mean = 0                                     degrees of freedom =       29

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Registration in the last three months

Original variable

```
      Registered in this |
sub-district in the last |
              3 months? |      Freq.     Percent        Cum.
------------------------+-----------------------------------
       1-Yes (completed) |         57       45.24       45.24
       2-Yes (under way) |         35       27.78       73.02
                   3-No |         11        8.73       81.75
4-Not yet, but scheduled |         23       18.25      100.00
------------------------+-----------------------------------
                   Total |        126      100.00
```

recoded as:

```
Regist3Months |      Freq.     Percent        Cum.
--------------+-----------------------------------
No or not yet |         34       26.98       26.98
    Under way |         35       27.78       54.76
    Completed |         57       45.24      100.00
--------------+-----------------------------------
        Total |        126      100.00
```

with 0 "No or not yet" I "Under way" 2 "Completed"

```
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
GammaMiss_Reg~t float  %9.0g                 Gamma recent registration
                                               vs. confid. ratings


                    GammaMiss_Regist
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.2645799      -.2645799
 5%    -.2494744      -.2494744
10%    -.1845281      -.1966643       Obs                  30
25%    -.1272115      -.1723919       Sum of Wgt.          30

50%    -.0538705                      Mean           -.0524799
                      Largest         Std. Dev.       .1049782
75%     .0341857       .0918023
90%     .0957123       .0996224       Variance        .0110204
95%     .1109833       .1109833       Skewness       -.1151999
99%     .1220204       .1220204       Kurtosis        2.220591

. ttest GammaMiss_Regist == 0
```

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
GammaM~t |      30   -.0524799    .0191663    .1049782   -.0916794   -.0132804
------------------------------------------------------------------------------
    mean = mean(GammaMiss_Regist)                                 t =  -2.7381
Ho: mean = 0                                     degrees of freedom =       29

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.0052         Pr(|T| > |t|) = 0.0104          Pr(T > t) = 0.9948
```

Recent fighting

```
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
fight30days     long   %19.0g      fight30days
                                               Contested area in the last 30 days


  Contested area in |
   the last 30 days |      Freq.     Percent        Cum.
--------------------+-----------------------------------
1-Frequent fighting |         49       38.89       38.89
2-Sporadic fighting |         52       41.27       80.16
      3-No fighting |         25       19.84      100.00
--------------------+-----------------------------------
              Total |        126      100.00

variable name   type   format      label      variable label
-------------------------------------------------------------------------------
GammaMiss_Fight float  %9.0g                  Gamma recent fighting vs. confid. ratings

                    GammaMiss_Fight
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.3894134      -.3894134
 5%     -.213264       -.213264
10%    -.1592126      -.1732789       Obs                  30
25%    -.0968165      -.1451464       Sum of Wgt.          30

50%    -.0229587                      Mean           -.007825
                      Largest         Std. Dev.       .1457089
75%     .0745147       .1350198
90%     .1654547       .1958895       Variance        .0212311
95%     .2772346       .2772346       Skewness        .0364375
99%     .3344305       .3344305       Kurtosis        3.788803

. ttest GammaMiss_Fight == 0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Gamma~ht |      30    -.007825    .0266027    .1457089   -.0622336    .0465835
------------------------------------------------------------------------------
    mean = mean(GammaMiss_Fight)                                  t =  -0.2941
Ho: mean = 0                                     degrees of freedom =       29

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.3854         Pr(|T| > |t|) = 0.7707          Pr(T > t) = 0.6146
```

REFERENCES

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

Anselin, L. (1995). "Local Indicators of Spatial Association - Lisa." Geographical Analysis 27(2): 93-115.

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

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

AWG (2013c). Syria Integrated Needs Assessment (SINA) [December 2013], Assessment Working Group for Northern Syria.

Benini, A. (2012)."A template for managing data in needs assessments, centered on sites, sectors, problems, and severity of needs [24 January 2012]." Geneva, Assessment Capacities Project (ACAPS). Retrieved 17 December 2014, from http://www.acaps.org/img/documents/data-management-note--datamanagement_note-1.pdf.

Hibbs-Pherson, K. and R. H. Pherson (2012). Critical thinking for strategic intelligence. Los Angeles, Cq Press, SAGE Publications.

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

Jeanty, P. W. (2010a). SPLAGVAR: Stata module to generate spatially lagged variables, construct the Moran Scatter plot, and calculate Moran's I statistics. Statistical Software Components, available from http://ideas.repec.org/c/boc/bocode/s457112.html, Boston College Department of Economics.

Jeanty, P. W. (2010b). SPWMATRIX: Stata module to generate, import, and export spatial weights. Statistical Software Components, available from http://ideas.repec.org/c/boc/bocode/s457111.htm, Boston College Department of Economics.

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

UNOCHA (2013). UNITED NATIONS DISASTER ASSESSMENT AND COORDINATION - UNDAC Field Handbook (6th edition). Geneva and New York, United Nations Office for the Coordination of Humanitarian Affairs.

UNOCHA (2014). Syrian Arab Republic - Governorates Profile (June 2014). New York and Geneva, United Nations Office for the Coordination of Humanitarian Affairs.

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

Wikipedia. (2014)."Goodman and Kruskal's gamma." Retrieved 15 December 2014, from http://en.wikipedia.org/wiki/Goodman_and_Kruskal%27s_gamma.

Wikipedia. (2015)."Moran's I." Retrieved 5 February 2015, from http://en.wikipedia.org/wiki/Moran%27s_I.

Wittek, R. and A. Zwitter (2014). From theory to analysis: H-AID methodology. Humanitarian Crises, Intervention and Security: A Framework for Evidence-Based Programming. L. Heyse, A. Zwitter, R. Wittek and J. Herman. London and New York, Routledge. 59-68.