Minimising Survey Measurement Errors in Complex Humanitarian Settings: Lessons Learned from the field

Presenting author: Ms. Nayana Das, Head of Research at IMPACT Initiatives[1]

European Survey Research Association Conference (Milan, July 2023)

Abstract

According to the 2023 Global Humanitarian Overview published by the United Nations' Office for the Coordination of Humanitarian Affairs (UN-OCHA), more than 300 million people around the world are currently in need of humanitarian assistance and protection, and addressing these needs would require over 50 billion USD in humanitarian aid funding. These needs are being further aggravated by global mega-trends such as forced displacement, armed conflict, food crises, climate change, and increasing poverty. In this global context where the scale and severity of humanitarian needs are constantly increasing, while the availability of resources to address these needs are not, the importance of robust data to inform effective planning and delivery of aid to crisis-affected populations has become more pertinent than ever.

To address this, several efforts have been made over the past few years to develop data production and analysis processes that promote a comprehensive understanding of the breadth and depth of humanitarian need around the world, including more streamlined roll-out of household surveys and other relevant research exercises. For more than ten years now, IMPACT Initiatives (IMPACT) has also been contributing towards these efforts by conducting research and analysis exercises to support aid actors planning and responding to a range of different humanitarian crises. In 2022 alone, IMPACT's teams collected primary data across more than 25 humanitarian crises globally; this data was collected directly from and about crisis affected-communities, through more than 200,000 household surveys, 260,000 key informant interviews and 1,500 focus group discussions.

Building on IMPACT's experiences and lessons learned from implementing surveys across multiple humanitarian crises, including both natural disaster and protracted conflict contexts, this paper will provide an analysis on the challenges and solutions for producing high quality survey data in such contexts. The paper seeks to answer the following two research questions:

1. What type of random and systematic measurement errors can impact production of high-quality survey data in humanitarian settings, and what are the sources of these errors?
2. What measures can be taken to ensure these errors are detected and addressed in a timely and systematic manner?

To answer these questions, this paper will provide an analysis of IMPACT's lessons learned from conducting surveys across different humanitarian contexts, including preliminary findings from two different methods that are currently being explored to strengthen the quality of survey data. This will be complemented by: 1) an in-depth review of existing survey methodological literature on measurement errors, and 2) findings from a structured online survey where IMPACT's research and data teams across different regions shared their perspectives on the types of measurement errors they commonly face, and some of the most effective ways to address them.

[1] Read more about IMPACT Initiatives here

1. Introduction: Survey research in humanitarian settings

Over the past decades, **survey research has emerged as the most frequently used method to collect data across a range of disciplines**, and almost any field of study which requires information on individual perspectives, experiences and behaviours relies on this method (Saris & Revilla, 2015). As Meyer et al. (2015) rightly note, "*large and nationally representative surveys are arguably among the most important innovations in social science research of the last century*", having also become a key source of information for official estimates of unemployment, poverty, and any other statistics needed to guide socio-economic policies around the world.

Within the humanitarian sector as well, survey research has proven to be a powerful tool to enable more accountable and evidence-based decision-making, with aid actors increasingly relying on survey data to understand the needs and vulnerabilities of crisis-affected communities. As with any other research process, the fundamental premise of surveys conducted in humanitarian contexts is to describe experiences of a target crisis-affected population, by observing a select sample within this population. In order to have an accurate and precise description of these experiences, the surveys thus need to be designed and implemented in the most robust way possible, especially to ensure that the sample selection minimizes random differences with the wider population. These surveys can then enable any actor involved in the design and implementation of aid programmes to determine the scale and severity of humanitarian need, and understand how this varies between different geographical areas (e.g. regions, districts, livelihoods zones) and population groups (e.g. displaced and non-displaced households).

1.1 Practical examples of survey methods in humanitarian settings

Although conducting robust survey research to generate an evidence base for humanitarian decision-making can be challenging, examples of high-quality, ethical, and actionable research do exist. Some of these are described below:

1. **Multiple Indicator Cluster Surveys (MICS):** A UNICEF-supported household survey programme launched during the mid-1990s, MICS monitors the situation of children and women around the world (Khan & Hancioglu, 2019). Over the years, MICS has become one of the world's largest household survey programmes, having covered more than 115 countries till date, and providing a wealth of data on topics such as fertility, mortality, unmet need, child development and nutrition (Khan & Hancioglu, 2019).[2]
2. **Multi Sector Needs Assessment (MSNA):** MSNA is a household survey whose primary objective is to provide crisis-wide data to inform more effective, context-appropriate delivery of humanitarian assistance, especially as part of the Humanitarian Programme Cycle.[3] Since 2016, IMPACT, through its flagship REACH Initiative, has led the implementation of MSNAs across 10-15 humanitarian crises on an annual basis. By collecting household-level multi-sectoral data from different geographical areas and population groups, the MSNA enables humanitarian actors within each crisis to better understand what is the prevalence and severity of needs within and across sectors; which groups and areas are most affected; what are the key drivers of these needs; and what is the co-occurrence of needs between sectors.[4]
3. **Comprehensive Food Security and Vulnerability Analysis (CFSVA):** CFSVA is a tool designed and used by the World

[2] Read more about MICS here: https://mics.unicef.org/
[3] Read more about humanitarian needs assessment and analysis here: https://www.unocha.org/themes/needs-assessment-and-analysis

[4] To learn more about the MSNA survey approach, see also: this blog post for the 2023 UN World Data Forum, and this article on the IMPACT website.

Food Programme (WFP) to "*understand and describe the profiles of food-insecure and vulnerable households, identify the root causes of hunger, and analyze the risks and emerging vulnerabilities among populations in crisis-prone countries.*"[5] A large household survey is the central element of this methodology, and this is meant to be complemented by desk review, qualitative community-level data, as well as a risk and response analysis.

4. **Standardised Monitoring and Assessment of Relief and Transitions (SMART) Surveys:** SMART is an inter-agency initiative launched by a network of humanitarian organisations and practitioners in 2002 to establish a systematized survey methodology that can provide "*critical, reliable information for decision-making*".[6] The SMART methodology is based on capturing data for two key health indicators to assess the magnitude and severity of a humanitarian crisis: nutritional status of children under-five, and mortality rate of the population.
5. **Demographic and Health Surveys Programme (DHS):** Established by the United States Agency for International Development (USAID) in 1984, the main objective of DHS is to improve the collection, analysis, and dissemination of population, health, and nutrition data and to facilitate use of this data for planning, policy-making and programme management across different countries.[7] Over the years, DHS has conducted more than 400 surveys in over 90 countries.
6. **Living Standards Measurement Study (LSMS):** As the World Bank's flagship household survey programme launched in the early 1980s, the main goal of LSMS is to "*foster the development and facilitate the adoption of new methods and standards in household data collection for evidence-based policymaking.*" [8] The key component is a multi-purpose survey that collects data on different dimensions of household and individual wellbeing, while also trying to understand the effects of various government policies on the living conditions of people in low and middle income countries.

1.2 Challenges with conducting survey research in humanitarian settings

While these examples do exist in practice, **the complex operational contexts associated with humanitarian settings introduces unique scientific challenges and conditions** that distinguish them from standard research practices. Because of these challenges, maintaining realistic expectations of data available in humanitarian settings, while pushing for more resources and methodological innovation to improve data quality, including close collaboration with local partners on the ground, remain essential (Fogarty International Centre, 2021).

There are three key challenges specific to survey research in humanitarian contexts:

- Firstly, due to **inaccessibility and time sensitivity factors that are unique to these contexts**, survey research in such settings need to navigate unusual barriers to implementation while trying to meet acceptable scientific standards (Guha-Sapir & Scales, 2020). While this is equally true for both post-disaster and protracted conflict contexts, the latter has additional challenges linked to resistance from parties to the conflict (including state and non-state armed groups), as well as safety and security challenges which brings elevated risks for both researchers and survey respondents (Guha-Sapir & Scales, 2020). Lessons learned from a research commissioned by Fogarty International Centre (2021) to inform health and

[5] Read more about CFSVA here: https://www.wfp.org/publications/comprehensive-food-security-and-vulnerability-analysis-cfsva-guidelines-first-edition

[6] Read more about the SMART methodology here: https://smartmethodology.org/

[7] Read more about the DHS programme here: https://dhsprogram.com/

[8] Read more about LSMS here: https://www.worldbank.org/en/programs/lsms/overview

nutrition interventions across conflict zones in Afghanistan, Mali, Pakistan and Somalia also found that security was "*a fundamental obstacle in all study contexts*", with restricted geographical access and safety concerns affecting sampling and data collection plans. From IMPACT's own experiences over the years, logistical and access constraints have led to increased challenges in the implementation of robust probability sampling techniques, especially for hidden, hard-to-reach populations (e.g. refugees outside formal camp settings, populations living in areas under occupation of non-state armed groups) and/ or populations on the move (e.g. refugees and migrants, internally displaced populations, returnees). Similarly, due to time pressures associated with survey research in humanitarian settings, especially for dynamic and rapidly evolving contexts, trade-offs often have to be made when designing methodologies to land on something "good enough" that can be delivered within the desired time-frame and with the resources and access available. Finally access and security considerations also creates barriers for direct oversight of the data collection exercise itself, with researchers often too far removed from the day-to-day survey implementation, thus limiting their ability to proactively identify, troubleshoot and mitigate issues as they arise.

- Secondly, researchers in humanitarian contexts also have to consider their data collection exercise with the **additional ethical responsibility of working with highly vulnerable and/or traumatized populations**, such as undocumented refugees and migrants, unaccompanied and separated minors, survivors of conflict-related violence, etc. For instance, when conducting a mixed methods research to assess mental health problems and barriers to accessing mental health care among refugees in urban areas of Turkey, Karadag et al. (2021) found that the high prevalence of past and present traumas required better communication skills and trust between respondents and interviewers, and often interviews had to be supported by senior researchers to decrease the risk of secondary traumatization. From its own experiences over the years, IMPACT has also found that design of survey methodologies sometimes has to compromise on the depth of information being collected, or even changing the methodology to something more appropriate, in order to ensure that the 'do no harm' principle is respected and data collectors or respondents are not exposed to any direct or indirect risk as a result of participation in the research; this includes the risk of re-traumatisation when asked to recall recent events and experiences for the survey. Moreover, assessment fatigue also needs to be a key consideration for the design of any new survey, since crisis-affected populations, especially in accessible areas, often tend to be overly assessed with different aid organisations conducting similar, albeit separate, research exercises within overlapping time-frames. Proper inter-agency coordination of data collection efforts is thus essential to ensure surveys are limited in scale and capturing only necessary 'need-to-know' information in a responsible, harmonized, and time-efficient manner.
- Finally, the **lack of reliable and up-to-date secondary data sources, especially standardized administrative methods for record keeping, data sharing and dissemination**, can also prove to be a key challenge for survey research in humanitarian settings. For instance, during their experiences conducting research in post-disaster humanitarian settings (including in Indonesia and India following the 2004 Indian Ocean Tsunami, the Philippines following Typhoon Haiyan in 2013, and Nepal following the 2015 earthquake), Guha-Sapir and Scales (2020) found this to be a key challenge for epidemiological studies trying to assess effects of the disaster on mortality, malnutrition, mental health, and diseases. While these issues are not unique to post-disaster contexts, the time-sensitive nature

of such research exacerbated the effects of these concerns and innovative methods had to be found to produce meaningful and context-appropriate data in such settings (Guha-Sapir & Scales, 2020). From IMPACT's experiences across different crises, lack of up-to-date, reliable population data in humanitarian contexts has also often led to challenges with the design of a comprehensive sampling frame, thus limiting researchers' ability to produce statistically representative data for different population groups of interest. Moreover, beyond administrative data, since humanitarian contexts tend to be data poor environments in general, this often means limited external, secondary data sources are available to triangulate and determine validity of collected data.

Interestingly, despite these unique challenges, **most of the survey literature has focused on understanding and mitigating errors within the fields of economics and marketing research, and very little is currently available for research in humanitarian settings,** with a few exceptions from public health and epidemiological studies. Building on IMPACT's experiences and lessons learned from implementing surveys across approximately thirty different humanitarian crises over the last decade, including both natural disaster and protracted conflict contexts, this paper aims to address this gap and provide an analysis of some of the key challenges and solutions for producing high quality survey data in such contexts.

The next section will provide a brief background overview of the definitions and concepts of survey measurement errors that will be used for the purpose of this paper. The following two sections then go on to discuss the most common types and sources of random and systematic measurement in humanitarian settings, and measures that can be used to detect and address these errors in a timely manner. The analysis presented in the paper is based on a mixed methods approach comprising of three components: 1) a review of IMPACT's internal lessons learned documentation from conducting surveys across different humanitarian contexts, including key findings from two different methods currently being explored to strengthen the quality of survey data; 2) an in-depth review of existing survey methodological literature on measurement errors; and 3) findings from a structured online survey where IMPACT's research and data teams across more than 20 countries shared their perspectives on the types of survey measurement errors they commonly face, and effective ways to address them. The online survey covered a total of 38 respondents from five different regions where IMPACT is currently operating.[9] However, since the survey used a non-probabilistic, snowball sampling strategy, findings should be considered indicative only.

2. Background: Definition and types of survey measurement errors

Evidence available from decades of survey research across different disciplines has shown that **survey data is not always reliable,** and even salient features of an individual's life such as years of schooling can be prone to error (Bound et al, 2001). Indeed, statistical analysis of 'error-contaminated data' dates back to the 1980s and early days of econometrics, but the topic remains fairly active even today (Schennach, 2016). Meyer et al. (2015) identify **three key issues increasingly affecting the quality of household survey data:** 1) unit non-response, i.e. when households do not want to / are not available to answer the survey; 2) item non-response, i.e. when households participate in the survey but do not want to / are not able to answer certain questions; and 3) measurement error i.e. when

[9] Of the 38 respondents, majority were from the Middle East and North Africa region (11), followed by Europe (8), and Sub-Saharan Africa (6). There were also 2 respondents each from Latin America and the Caribbean, and Asia (excluding Middle East). Data collection was conducted between 27th June – 6th July 2023. Finally, the survey also included 9 respondents from IMPACT's global support teams in Geneva, Switzerland.

households provide answers to the survey but the answers themselves are not accurate. They also note that survey quality is in decline in general since "*households are overburdened by surveys leading to a decline in many measures of survey cooperation and quality*".

In simple terms, **'measurement error' in survey research is the difference between the *true* value and the *observed* value of a specific phenomenon.** It is sometimes also referred to as observation error, experimental error or 'total survey error' i.e. the deviation of a survey response from its underlying true value (Biemer, 2010). The reliability of a survey is thus a function of its total survey error, which is expressed as a function of the difference between the overall population's mean true value and the mean observed value obtained from the respondents of a particular sample (Assael & Keon, 1982).[10] Such measurement errors have been studied extensively in survey literature over the years, and **can be classified into two categories: random and systematic** (Bhandari, 2021).

Random error, also sometimes referred to as classical error, occurs when the difference between the true and observed value occurs by chance, usually because of imprecise or unreliable measurement instruments or poorly controlled experimental conditions (Bhandari, 2021). For the former, an example could be faulty weighing scales used to record observations or incorrectly calculated variables within the dataset due to a technical glitch in the mobile data collection software. For the latter, random error typically arises when the sample selected is not a perfect representation of the population of interest (Assael & Keon, 1982). This is why random error is also sometimes referred to as sampling error i.e. if by chance there is an over or under-representation of a certain sub-group of the population, such as individuals who do not work and therefore tend to be at home at the time of the household visit (Fowler, 2009).

This type of error mostly affects the *precision* of data collected i.e. the extent to which the same measurement will be reproducible under similar circumstances. As such, random errors might not be as problematic, especially if data is being collected from a sufficiently large sample size, since the errors in different directions might cancel each other out when calculating summary statistics (Bhandari, 2021).

Systematic error, also sometimes referred to as non-sampling error, occurs when there is a consistent and standardized difference between the true and observed value because of an underlying bias within the survey design or data collection process. Some common examples of such an error include inaccurate responses due to response biases (including social desirability bias), enumerator fatigue, and sampling biases. Systematic errors can also sometimes be due to non-response biases (Assael & Keon, 1982).

Unlike random error, a systematic error affects the *accuracy* of data collected i.e. how close is the *observed* value to the *true* value (Bhandari, 2021). Systematic errors can thus be much more problematic for the analysis, because they can skew the data in unknown directions and potentially lead to false conclusions.

Overall, the impact of measurement error on survey findings depends on the magnitude of the error relative to the true variation, as well as the joint distribution of the measurement errors and true variables (Bound et al., 2001). Several studies in the past have shown that measurement errors in surveys can have considerable effects on the results obtained, with as high as 50% of the variance of observed variables in survey research being due to such errors (Saris & Revilla, 2015). Needless to say, both random and systematic errors can be quite problematic and it is important to identify and mitigate them as much as possible to ensure inferences derived from survey data are as precise and accurate

[10] Total survey error is measured as the mean squared error of the mean sample response around the population mean true value (Assael & Keon, 1982).

as possible (Biemer, 2010). The following two sections will discuss this in more detail.

3. Most common types and sources of survey measurement errors in humanitarian settings

As noted above, raw survey data is imperfect and analysing such data to draw accurate conclusions requires an understanding of its most significant shortcomings (Bound et al., 2001). Over the years, research on measurement errors across various disciplines has provided a rich empirical foundation for understanding under what circumstances survey responses are most likely to be subject to error (Bound et al., 2001). In general, **survey measurement errors can arise from a variety of different sources,** including sampling design, specific characteristics of the population of interest, topic(s) covered in the survey, design of the questionnaire, and overall conditions of data collection (Alwin, 1989). However, the source of the error and measures to address it will vary, depending on whether the error itself is random or systematic.

Within humanitarian contexts specifically, both random and systematic errors can lead to challenges with ensuring quality of survey data. Indeed, when asked about their perceptions of the most common types of errors in humanitarian contexts, majority of respondents from IMPACT's research and data teams across different countries (21/38) reported that both can be equally common, depending on the data collection context. However, a higher number of respondents (10) perceived systematic errors to be more common, compared to random errors (4).

3.1 Examples and sources of *random* errors in humanitarian settings

The main source of random error in any survey research is sampling error, which is introduced when there is a chance variation, usually in an unknown direction, between the characteristics of the sample and the true characteristics of the target population (Fowler, 2009). For instance, it is often the case that unemployed men and/ or women are over-represented in household surveys, simply because data collection takes place during working hours (Karadag et al., 2021).

Aside from chance variations, sampling error can also be directly due to a frame error, which arises when the sampling frame leads to omissions or non-coverage of specific parts of the population (Biemer, 2010). To take a concrete example, when IMPACT facilitated the first MSNA in Northeast Nigeria in 2019, sampling coverage in the most conflict-affected state – Borno – was limited to secure and accessible urban centres only. When the survey findings were being reviewed, interestingly, water, sanitation and hygiene (WASH) needs were found to be more severe in Ademawa and Yobe states, even though Borno is known to be more in need overall. Upon further investigation, it was found that this was due to a frame error and gaps in sampling coverage, since compared to the other two states, sampling coverage in Borno was limited to urban centers which also tend to receive a lot of humanitarian assistance, including for WASH services. Meanwhile, the coverage was more comprehensive in the other two states, including rural areas where access to proper WASH services and facilities has been historically challenging due to chronic under-development.

In humanitarian settings where accurate, up to-date data on the population of interest is rarely available, frame error can thus be a key source of random errors, since **inability to define a proper sampling frame is a key challenge, which subsequently complicates the implementation of representative sampling methods** (Guha-Sapir & Scales, 2020). Meyer et al. (2015) rightly note that "*Coverage error could explain some of the significant underreporting we find if the sampling frame for the surveys we examine (typically based on the non-institutionalised Census population) does not capture the entire population*". Even though two-stage cluster

sampling is sometimes used to circumvent the lack of comprehensive household lists, this also comes with its weaknesses, especially in terms of keeping design effects within acceptable limits. In post-disaster settings where surveys are trying to measure impact on public health, this can be especially problematic since morbidity and mortality often tends to occur in clusters, meaning that design effects can become inflated, thereby weakening the overall quality of survey results (Guha-Sapir & Scales, 2020).

Moreover, population sampling in conflict and post-disaster settings, especially the ability to ensure fully randomised selection of respondents, is further complicated by: 1) **large-scale displacement** with a large part of the population of interest constantly on-the-move, and 2) **limited access to the population of interest** due to destruction of roads and infrastructure, and/ or security considerations. For instance, when trying to conduct surveys with "hidden" populations like refugees and migrants in urban contexts, or specific demographic groups like internally displaced women, IMPACT has faced challenges with designing and implementing rigorous probability sampling strategies. As a result, teams on the ground have often had to resort to alternative methodologies, including qualitative research or the use of non-probability strategies such as quota and snowball sampling, as appropriate.[11]

Similarly, when conducting a mixed methods research to assess mental health problems and barriers to accessing required care among refugees in urban areas of Turkey, Karadag et al (2021) found that research for refugees in urban settings pose different challenges than research in camps. Specifically, obtaining a representative sample was not feasible due to difficulties with contacting the population of interest, which in turn was due to lack of publicly available demographic data on a district level, presence of undocumented asylum seekers, as well as high mobility among urban refugees.

Aside from sampling errors, **another source of random error could be conditions of the survey measurement process,** both in terms of the mode of the questionnaire i.e. self-administered or interviewer-administered (Alwin, 1989), as well as the setting or environment within which data collection is being conducted (Biemer, 2010). Specifically, this can result in chance variations between the "real world" and the data collection environment, because of which observations from the sample may not match realities of the wider population.

Further confirming all of the above, almost half of the online survey respondents (17/38) from IMPACT's research and data teams across different countries stated that "*Inaccurate population data to build a robust sampling frame for the population of interest (e.g. for locating IDPs out-of-camps, returnees, etc.)*" was the most common source of random errors for surveys implemented in their respective data collection contexts. Additionally, quite a few respondents (16/38) also perceived "*Chance variations in real world and experimental contexts (e.g. over-representation of unemployed household members due to time of data collection)*" as another common source of random errors.

3.2 Examples and sources of *systematic* errors in humanitarian settings

Unlike random error, the **sources of systematic errors can be more complex and varied, and are typically linked to two factors: non-response error and response error** (Assael & Keon, 1982). While non-response error occurs when some sample members do not respond to the survey (thus making responses an unreliable representation of the population), response errors occur when

[11] As an example, see this research Terms of Reference (ToR) with methodology description for an annual MSNA conducted with refugees and migrants in Libya, and this research ToR conducted on access to livelihoods for displaced women in Iraq.

the sample provides inaccurate responses to the survey questions (Assael & Keon, 1982).

According to Meyer et al. (2015), unit non-response (i.e. when a household in a sampling frame is not interviewed at all) has become an increasing concern for systematic errors over the years. For instance, unit non-response rates reportedly increased by 3-12% over the 1990s for six U.S. Census Bureau surveys. Based on data recorded for past household surveys, the most common reasons for unit non-response include inability to reach the household (e.g. due to faulty phone network connection), refusals due to lack of interest, time and / or motivation, as well as privacy concerns (Meyer at al., 2015). Moreover, even if a household agrees to participate, it can choose to not respond to specific questions, thus resulting in item non-response which can also introduce systematic error in the survey.

Meanwhile, **response errors can be due to several reasons** such as purposeful mis-reporting (for e.g. due to social desirability bias), faulty recall, enumerator and/ or respondent fatigue, un-favourable interview conditions, poor questionnaire design and so on. The rest of this section will look more closely at **examples and sources of some of the most commonly encountered sources of response errors in humanitarian contexts:**

1. **Questionnaire design:** The design of the questionnaire and encoding of information determines *a priori* the quality of data that will be produced through the survey. Poorly designed questionnaires, including issues with coding of the questionnaire for mobile data collection software, can thus be a key source of systematic response errors. An underlying factor for this is specification error i.e. when the concept implied by the survey question differs from the concept that should have been measured in the survey (Biemer, 2010). As a result of this, the wrong construct or concept ends up being measured, and eventually estimated, by the survey. Since the 1980s, several studies have looked at the effects that wording of survey questions can have on their responses and unsurprisingly, almost all of these studies conclude that formulation of the questions can have a considerate effect on the results obtained (Saris & Revilla, 2015). In addition to construct and wording, IMPACT's lessons learned have also shown that avoiding unnecessarily long questionnaires, as well as proper translation of the questionnaire to the relevant local language(s) for each area and population group, are also important ways of ensuring more accurate responses.
2. **Content of the survey:** The types of topic(s) being covered by a survey can also be a key source of systematic response errors. While there is more limited room for systematic errors within surveys covering factual content, which would be objective information regarding the respondent or his/ her household,[12] surveys measuring more subjective content such as beliefs, perceptions and attitudes could be less reliable because they require subjective assessments of specific experiences (Alwin, 1989).[13]
3. **Cognitive processes:** Past research on survey methodologies have categorized the survey question and answer process as a four-step process which involves: 1) understanding of the question, 2) retrieval of information from memory, 3) assessment of the correspondence between the retrieved information and the requested information, and 4) effective communication. Most of the literature on measurement error focuses on the third step, and classifies the under or over reporting of certain events or behaviours as the result of retrieval failure on behalf of the respondent, often due to the length of the recall period i.e. the longer the recall period, the greater the expected bias due to retrieval and reporting errors (Bound et al., 2001). In addition to this,

[12] Examples of factual content include respondent's demographic characteristics, household size, demographic break-down of the household, etc.

[13] Examples of subjective survey content include political preferences, barriers to accessing services, perceived safety risks for household members, etc.

misunderstanding of the question and not having the information needed to answer but answering anyway could also be a key source of response errors (Fowler, 2009).

4. **Social desirability bias:** Another specific cognitive process that can lead to systematic response errors is social desirability bias, i.e. distorting certain answers to look good. In other words, even if the respondent is able to retrieve accurate information concerning certain events or behaviours, he or she may deliberately choose to edit this information due to social desirability bias (Bound et al., 2001). Past research has shown that systematic errors linked to social desirability bias typically occur when surveys ask about socially and personally sensitive topics which are likely to elicit patterns of underreporting for socially undesirable behaviours and attitudes, or overreporting for socially desirable behaviours and attitudes. Some examples of this from humanitarian survey research could be under-reporting of safety and security concerns for female household members, under-reporting child protection issues such as child labour and marriage, over-reporting household income and expenditure levels, and over-reporting food security situation of the household. Indeed, quite a few online survey respondents (10/38) from IMPACT's research and data teams across different countries, especially in the Middle East and North Africa region (5), perceived "*Response bias or social desirability bias e.g. participants answer incorrectly on purpose or are prompted to answer in a certain way*" to be the most common source of systematic error for surveys implemented in their respective data collection contexts.
5. **Enumerator fatigue and / or demotivation:** Aside from errors on the side of the respondent, response errors can also arise if the enumerators themselves are fatigued or demotivated for the data collection exercise. This could lead to response errors either due to intentional data falsification (e.g. enumerator answering questions themselves without actually conducting the survey) or basic data entry mistakes (e.g. enumerator is rushing through the questionnaire to complete the survey as quickly as possible). The risk of response errors due to enumerator fatigue is especially high in humanitarian contexts when access limitations makes direct oversight of data collection challenging, and researchers have to remotely train and supervise data collection teams from afar. Further confirming this, seven online survey respondents from IMPACT's research and data teams across different countries perceived "*Experimenter drift or enumerator fatigue e.g. enumerators become fatigued, bored or de-motivated after long periods of data collection and start to drift from standardised procedures, including sampling techniques*" to be the most common source of systematic error for surveys implemented in their respective data collection contexts.
6. **Sampling biases:** In humanitarian contexts where security and access challenges often lead to known and intentional biases in sample design, sampling biases can also be a key source of both response and non-response errors. Some common examples of this include when survey responses are exclusively collected from male heads-of-households, often from older age groups, and populations in hard-to-reach locations, including rural areas and informal camp settings, are systematically excluded from the sampling frame. Indeed, almost half of the online survey respondents (17/38) from IMPACT's research and data teams across different countries perceived "*Sampling bias"* to be the most common source of systematic error for surveys implemented in their respective contexts.
7. **Data collection conditions:** Finally, like random errors, systematic measurement errors can also be introduced by certain conditions of the survey measurement process. According to Bound et al. (2001), specific features of survey conditions that could lead to systematic errors include: the mode of

data collection (i.e. in-person, phone, online or self-administered), characteristics of the interviewer (e.g. sex, age, ethnicity, etc.), type of data collection (i.e. cross-sectional or longitudinal), and the source of data collection (i.e. the data collection organization). For example, a study conducted by Assael and Keon in 1982 found that although both phone and in-person interviews require immediate responses, the involvement in an in-person interview may lead respondents to concentrate harder and put more effort into giving accurate responses. IMPACT's lessons learned across different contexts has also shown that cultural dynamics, especially trust between the data collector and respondent, can also have an impact on the accuracy of responses. For example, in some cultural contexts, a female household member might not feel comfortable discussing specific protection concerns for women and girls with a male data collector, thus leading to potential under-reporting of such issues. In 2019, IMPACT's team working on the Rohingya refugee response in Cox's Bazar district of Bangladesh also conducted a pilot study to determine the feasibility of using Rohingya enumerators for survey data collection processes, and to understand how data collected by Rohingya enumerators may vary in response and quality compared to data collected by Bangladeshi enumerators. The study provided some interesting insights on the consistency of responses and inputs based on the type of enumerator. Specifically, it was observed that households did provide different responses when asked sensitive or perception-based questions depending on the background of enumerator; for example, when asked about safety and security conditions for households in the camps, a higher proportion of households reported negative perceptions to Rohingya enumerators (13%) compared to Bangladeshi enumerators (0%).[14]

[14] See also: IMPACT Initiatives (2019), *Participation of Rohingya Enumerators in Data Collection Activities: Findings from a Pilot Assessment in Cox's Bazar, Bangladesh.* https://rb.gy/vb121

4. Effective measures to minimise survey measurement errors in humanitarian settings

While there is quite a bit of evidence available within survey methodological research on the types of random and systematic errors, including differential effects of the varied sources of these errors, empirical evidence is limited and inconsistent on the direction and magnitude of the error caused by each of these sources (Bound et al., 2001). Additionally, while a large part of the literature looks at the problem of "*recovering true error-free quantities from error-contaminated data*", this is based on an underlying assumption that the distribution of the error itself is known, which is not always the case (Schennach, 2016).

All of the above makes it challenging for researchers, especially in complex humanitarian settings, to systematically identify and address measurement errors. This section will explore this in more detail, trying to provide some insight specifically for humanitarian research contexts on: 1) key overall considerations for ensuring survey data quality, 2) some established methods to address measurement errors and the feasibility of their application in humanitarian contexts, and 3) two specific methods being explored by IMPACT to improve survey data quality across different contexts.

4.1 Some overall considerations for ensuring survey data quality in humanitarian settings

Evidence available within survey methodological literature suggests that addressing random errors can be straightforward, as these types of errors can be controlled by careful selection of the sample and ensuring large enough sample

sizes (Assael & Keon, 1982). On the contrary, **minimising systematic, non-sampling error tends to be more challenging and harder to control** (Assael & Keon, 1982). From some of their own past research experiences, Bound et al. (2001) concluded that non-classical, systematic measurement errors need to be taken more seriously, both in terms of assessing the likely biases in analysis that don't take any account of measurement error, as well as in devising techniques that can minimize and correct for such errors.

Regardless of the type of error, **the first step in ensuring quality of survey data is to identify the sources of errors in a given data collection context** (Biemer, 2010). In order to do this, a robust data cleaning process needs to be followed for each survey, throughout the data collection process. This is especially important for the complex operational data collection environments that are characteristic of humanitarian contexts, where as a result of access and security barriers, researchers often have to supervise data collection remotely and have limited control over the implementation of the survey on the field.

According to IMPACT's Data Cleaning Guidelines for Quantitative Research (available here), there are a **few steps that need to be followed to implement data cleaning before, during and after data collection:**

- **Before data collection:** First and foremost, all IMPACT teams need to ensure questionnaires and related tools are well designed, and also set up a clear process, with SOPs, for data cleaning and management of data collection. At minimum, these SOPs should clarify: 1) division of responsibilities for each team member during the data checking and cleaning process; 2) a breakdown of the data collection plan, clarifying how data collection will be tracked, procedure to be followed when intended respondents are not found in expected locations, and procedure to be followed in the case of both unit and item non-response; 3) summary of steps for the data checking process i.e. what type of checks will be applied and why, as well as how to execute verifications, corrections and deletions within the data; and finally 4) a detailed overview of pre-defined thresholds or 'red flags' to identify and address potentially problematic data with serious accuracy concerns. In addition, all teams are also required to pilot and test each data collection tool and sampling methodology, and provide all necessary trainings for data collection teams (including enumerators and field officers supervising data collection). Biemer (2010) also notes that more systematic enumerator training, even if costly and time consuming, is extremely necessary to minimize serious interviewing errors. Similarly, in a 2017 guidance note published by WFP's Regional Bureau in Egypt (see here), highlights 'Tool design', including pilot testing of the tool before data collection, as the first step for ensuring quality of survey data. When asked in the online survey what they have found to be the most effective way to minimize random and systematic errors across their respective contexts, "*properly designed questionnaires*", "*pilot and testing of data collection tool and methodology*" and "*carefully control data collection tools and conditions*" was commonly mentioned by quite a few representatives from IMPACT's research and data teams across different countries. Pilot and testing was perceived to be an especially effective measure by respondents in sub-Saharan Africa and Europe.
- **During data collection:** Once data collection has begun, all teams are required to supervise data collection and monitor, check and clean incoming data on a regular (preferably daily) basis, while maintaining a consistent feedback loop with data collection teams throughout to ensure all identified issues are being followed up on as data collection progresses. Moreover, the data team is required to maintain a consistent record of all checks done and actions taken during data cleaning, while ensuring that the

unedited raw data version is always preserved for transparency.

- **After data collection:** Once data collection is complete, teams finalise data cleaning and submit the final, cleaned dataset with all accompanying documentation to IMPACT's global research team for a final review. The global team is responsible for validating the data and data cleaning process documentation, in order to ensure that the required minimum standards have been fulfilled and the dataset is ready for analysis and external dissemination. Additionally, at this stage the research team in country is required to review and consolidate overall lessons learned, to be incorporated for future data collection exercises.

Confirming the importance of the above steps, when online survey respondents from IMPACT's research and data teams were asked to provide specific examples of an effective method that they have applied to minimise measurement errors in recent surveys, most respondents provided examples related to setting up of rigorous data collection and cleaning processes. This includes 1) training and re-training of field officers and enumerators; 2) pilot and testing of tools and methodology prior to data collection; 3) drafting comprehensive data cleaning SOPs and ensuring their implementation on a daily basis; and 4) use of automated tools and nuanced methodologies to improve rigour and efficiences of data quality checks, especially for more complex indicators (e.g. food security outcomes). Finally, the importance of well-designed questionnaires was also reported as an effective measure for response errors, especially in terms of coding and definitions.

As a key reference guide for the different data cleaning steps outlined above, IMPACT's guidelines also include a list of **'minimum standards' which all survey datasets are required to fulfil.** These 'minimum standards' **can be summarized in four categories:**

1. **Survey metadata:** Key checks here include removing any duplicated records (i.e. all records should have a unique identifier), and ensuring incoming data is consistent with the intended sampling strategy (e.g. number of household records per district is in line with the stratification plan).
2. **Data protection:** The main requirement here is to ensure that any information that could be used to identify individuals (respondents or data collectors) or households is removed from the dataset prior to any further internal or external dissemination. This includes both direct identifiers (e.g. individual name and contact details, household geo-location, etc.) as well as indirect identifiers (e.g. a combination of household size, head of household's name, and village name).[15]
3. **Enumerator metadata:** The main objective here is to monitor survey patterns and enumerator behaviour in order to identify data entry errors or potential data falsification. Key checks include ensuring enumerator's interview speed (i.e. time taken for the survey) is reasonable, and ensuring that none of the enumerators consistently follow the shortest questionnaire path or exact same path i.e. providing same responses across multiple records.
4. **Logical checks:** This includes both vertical checks to ensure there are no inexplicable or impossible outliers (i.e. an observation/ a specific data point that lies an abnormal distance from other values) within the relevant variables, as well as horizontal checks to ensure there is logical coherence between the different connected responses within each survey record. This is especially relevant to minimize systematic errors in larger surveys like the MSNA, where many different questions are asked for which responses do need to

[15] IMPACT has a separate guidance note available on how to manage personally identifiable information during each research process (available here). In addition, the Inter-agency Standing Committee has also recently published a broader, more system-wide 'Operational Guidance on Data Responsibility in Humanitarian Action' which can be accessed here.

speak to each other. One example of logical incoherence from an MSNA conducted by IMPACT in 2019 in Bangladesh was when several households were reporting to be satisfied with the latrines they were using, while also reporting that overflowing of latrines was the main issue they were facing with their sanitation facility. Similarly, for an MSNA survey implemented in Burkina Faso in 2022, one of the logical inconsistencies that were checked was if households reported having "*no barriers*" to accessing healthcare, but also reported they did not have access to healthcare facilities nearby. Inconsistencies like these need to be carefully checked and understood, and data points cleaned as needed.

Similar to these guidelines used within IMPACT, MICS also follows a multi-pronged approach to data quality assurance (Khan & Hancioglu, 2019). This includes: 1) extensive training of enumerators for an average of three to four weeks prior to data collection; 2) organizing enumerators into teams, including a supervisor who coordinates data collection activities and re-visits select households for quality control; 3) transmitting and storing data to a cloud server on a daily basis, which is then checked by a 'survey manager' every week with tables measuring data quality indicators disaggregated by team and interviewer; and 4) taking immediate corrective action as issues are identified while enumerators are still in the field, including pausing data collection for certain teams and re-training or re-recruiting if necessary.

Now that we have an overview of how data cleaning checks and processes should be set up to ensure survey errors are detected in a timely manner, the remainder of this section will discuss some specific measures that can be used to address: 1) random or systematic errors specifically linked to sampling issues, and 2) systematic response errors.

4.2 Measures to address random or systematic errors related to sampling

As previously noted in section 3, issues related to sampling can be a key source of both random and systematic measurement errors. For the former, **an effective first measure to control for this is to ensure careful selection of the sample with robust randomisation techniques and increase sample sizes as required** (Assael & Keon, 1982). For instance, one approach that IMPACT systematically tries to apply across almost all its surveys is to ensure that some degree of stratification is always included within each sampling design. By doing so, we not only manage to obtain large (aggregated) samples for each survey, but also ensure that varied experiences of different crisis-affected population groups, including displaced and non-displaced households across different geographical areas, are adequately accounted for.

Meanwhile, when simple random sampling techniques for household selection cannot be used due to missing household lists, **IMPACT has also deployed more nuanced approaches, including GIS-based sampling, to ensure data collection teams are indeed able to randomize household selection in the field.** In very simple terms, this approach involves using GIS software to generate and distribute randomised GPS points on a map covering the area of interest. The distribution of GPS points is weighted by population density, should this vary across the targeted area, as indicated by available spatial data. These points are then provided to each enumerator via a mobile navigation software (e.g. maps.me) on their hand-held data collection devices, and they are required to locate and survey a unit (i.e. household) closest to each point, usually within a pre-defined buffer distance as relevant to the context. The added value of this approach is that randomised household selection can be verified, simply by comparing the assigned GPS point with the GPS point recorded in the survey for each assessed household. Having said that, in order for this approach to work well, certain pre-requisites need to be in place, including: 1) availability of accurate, up-to-date shape files for administrative boundaries of the areas of interest; 2) availability of reliable

spatial or other data indicating the distribution of the population and population density across the targeted area; and 3) well-trained and skilled data collection teams that have the capacity to use maps.me or similar navigation software to locate sampled GPS points on the ground.[16] Because of this, IMPACT has found that while this technique can work well in some contexts, especially in more stable protracted conflict contexts, it is more challenging to implement in some other more dynamic and complex contexts (e.g. South Sudan, Central African Republic, Yemen), as well as for specific population groups such as returnees, female-headed households, and refugees or migrants outside formal camps.

In general, in humanitarian contexts where accurate population data to design sampling frames is rarely available, and sampling errors are unavoidable due to access and security considerations, addressing both random frame errors as well as systematic response and non-response errors becomes much more challenging. In these instances, **more creative solutions need to be considered, including the use of mixed methods with qualitative approaches and/ or geo-spatial analysis, as appropriate.** For example, in Afghanistan where no official census has been conducted since the 1970s, and there is no complete or coherent list available of all existing population settlements across the country, IMPACT's team in this context regularly face challenges linked to frame errors, especially for sampling and collecting data outside large urban areas. To address one such issue, the team used satellite imagery to look at infrastructural features and draw urban boundaries around the provincial capitals, which are considered to distinguish themselves from other areas in their geographic dimensions, population density, relationship to their province and region, and city functions/ characteristics. Furthermore, a list of approximately 50,000 settlements and their GPS points have been compiled till date through all past and on-going research exercises, and these are then analysed by the GIS team to map each of their boundaries.

Similarly, from their past research experiences in post-disaster contexts, Guha-Sapir and Scales (2010) concluded that "*weaknesses attributable to faulty sampling or data collection in emergent situations can greatly benefit from qualitative techniques to elucidate the findings"*. For instance, when conducting a study to understand the risk factors for mortality, injury and epidemic-prone diseases following the 2004 Indian Ocean tsunami in Indonesia and India, they found that only by working with local partners and gathering age and sex data for most deaths were they able to draw a systematic study sample and provide a statistically robust design for the study. However, they also found that relying only on statistical analyses would be misleading, especially in situations where sampling was weak, and focus group discussions were organized to contextualise quantitative findings, and provide stronger findings.

Another specific measure that can help address sampling-related errors includes mitigation of unit non-response. Till date, several methods have been proposed and used by survey researchers to improve unit non-response in surveys, including advance notification of the survey via email or text message, increasing the number of the times the potential respondent is contacted, strengthening training of interviewers, and offering financial incentives for participation (Meyer et al., 2015). However, even if such efforts can increase response rates, they do not necessarily lead to a reduction in systematic errors and responses biases, and sometimes can make the biases even worse if they encourage groups that are already over-represented in the survey (Meyer et al., 2015). Of course, if non-response arises randomly across the population, survey data could still lead to unbiased estimates of distributions. However, exploring whether unit nonresponse is random can be difficult because researchers

[16] For a more detailed overview of IMPACT's sampling approaches, including a step-by-step guide on the implementation of this GIS-based sampling approach for host community surveys in Jordan, please refer to IMPACT's Research Design Guidelines (2020) available here.

typically have only limited information on the characteristics of non-respondents. This makes it difficult to determine the extent to which unit non-response leads to sampling or response bias by survey and question, and while there are examples of substantial bias in some cases, in other cases the resulting bias can be small or easily mitigated by appropriate weighting (Meyer et al., 2015).

Finally, **another important measure to address sampling-related errors, analyse the overall reliability of survey measurement and avoid biased statistical estimates, is the use of variance estimation techniques during analysis** (Alwin, 1989). But this estimation of reliability is not always straightforward, and especially difficult for cross-sectional surveys (Alwin, 1989). In general, while certain survey variance estimation techniques have been developed over the years (for example, replicated sampling, balanced repeated replication, jackknife-repeated replication, the bootstrap method, and the Taylor series method), the feasibility and relevance of their application in humanitarian contexts need to be further explored. Moreover, such techniques can be more useful for correcting random measurement errors, not necessarily systematic measurement error (Bound et al., 2001). In collaboration with Statisticians without Borders (SWB), IMPACT is currently exploring ways to determine extent of random errors by analysing design effects for cluster sampling survey designs; more explanation and preliminary findings from this is presented in sub-section 4.4 below. Meanwhile, the following sub-section discusses in more detail how triangulation measures, including validation studies, can be useful to detect and address systematic errors caused by response biases or other related sources.

4.3 Triangulation measures and validation studies to address systematic response errors

Since the 1980s, in disciplines like econometrics and labour economics, efforts have been made to conduct validation studies to verify accuracy of reported survey data, specifically by comparing survey responses and administrative data on the same variables (Kapteyn & Ypma, 2007). For example, as early as the 1980s, Assael & Keon (1982) looked at actual telephone usage data and compared this to reported usage data from survey responses. By obtaining actual and reported data for three survey questions, they were able to compare survey designs and sources of survey errors to conclude that random sampling error was only a minor contributor compared to systematic error. In the field of economics, approaches like the employee-employer survey method, which involves comparing workers' self-reports on topics like earnings and benefits against reports of their employers, have also proven useful to an extent to determine the nature and implications of errors within such surveys (Duncan & Hill, 1985). However, a key limitation is the unknown level of error in the validation source i.e. employer reports.

In general, the added value of such validation studies is that they can give direct evidence on the nature of the measurement error by allowing comparison of survey responses to 'true' values of the same variable, with the latter often obtained from employer or administrative records (Bound et al., 2001). The measurement of the error is typically expressed as a mean and some measure of dispersion, and correlations between the two relevant measures are reported i.e. validity of that measure is the correlation between the measure and the underlying construct that the measure is trying to capture (Bound et al., 2001).

Validation data to apply such an approach can be obtained in two ways: data collected as part of the same primary data collection effort (also sometimes referred to as internal multiple indicator data), or data coming from external (often independent) studies such as employer records or administrative data (Bound et al, 2001). The former, although more time and resource intensive, is preferred and is more likely to yield meaningful results (Bound

et al., 2001). Having said that, the latter can also prove to be useful especially for reverse record checks; for instance, if the measure of interest is a discrete event (such as hospitalization, industrial accident related to a particular job), reverse record checks which involves sampling and interviewing select respondents from the original administrative (or validation) records to then confirm certain reported behaviours, can prove quite effective to verify under-reported issues (Bound et al., 2001). Meanwhile, prospective record checks, which involve first interviewing the survey sample and then verifying reported behaviours with administrative records, are better placed to verify over-reporting of specific events (Bound et al., 2001). The most effective way to verify both under and over-reporting is doing a complete record check if all of the relevant records can be identified, but this is quite rare since it requires the implementation of probability sampling for all units of the population of interest and the availability of validation information for all sampled units (Bound et al., 2001). The relevance of validation studies can also be further enhanced if data can be collected for a random portion of the same survey, rather than from a separate external source and/ or another convenience or non-probabilistic samples (Bound et al., 2001), but this can be quite time intensive and lead to respondent fatigue.

Another nuanced validation approach includes re-interview designs like the 'test re-test' approach, which some of the survey methodological literature also recommends as an effective measure to improve reliability of survey data. This approach, which was first introduced by Lord and Novick in 1968, aims to estimate the quality and reliability of questions by measuring the same concept in two different ways, within the same survey (Biemer, 2010). For example, as part of a Digital Needs Assessment Toolkit developed by IMPACT in 2020, to confirm if respondents knew how to obtain additional credit for their sim cards, the following three questions were asked: 1) Do you know how to recharge prepaid credit? 2) Do you know how to top up airtime? 3) Do you know how to add account balance? In addition to identifying the "optimal" way to ask this question by selecting the version most frequently selected by respondents as "easiest to understand", differences in the responses to each version were also investigated to identify misunderstandings, if any. However, this also means that the number of questions within the survey will be at least twice the number of concepts to be measured, thus making this a costly, time-consuming and often unfeasible procedure to apply (Saris & Revilla, 2015). In humanitarian contexts where survey fatigue among crisis-affected populations already tends to be quite high, this could also be an un-ethical ask of the respondents' time.

Overall, if implemented successfully, validation studies can enable researchers to understand bounds of the bias within survey data, even if not eliminate them. In other words, they allow researchers to "*assess the magnitude of measurement errors in survey data, and the validity of the classical assumption*" (Bound et al., 2001). Validation data thus enables identification of specific parameter estimates in the presence of (often arbitrary) patterns of measurement error, and their value is further enhanced if they include not just data on the key variables being validated but also on other variables that could be used in conjunction with those variables (Bound et al., 2001).

Over the past years, IMPACT has explored the use of such validation studies, albeit in slightly nuanced ways, to understand the extent of survey measurement error across different contexts. Two such examples are discussed in more detail below:

- **Household survey validation with dual respondent data – Afghanistan, 2021:** As part of a nationwide household humanitarian needs assessment survey (MSNA) conducted in 2021, for specific variables and in a select number of provinces, two sets of responses were collected from male and female

respondents within the same household.[17] The objective of this dual-respondent methodology was to ensure female voices were adequately captured within the survey, especially for areas like livelihoods and protection where experiences of female household-members have a likelihood of being mis-represented by male heads of households given the cultural context of Afghanistan. Additionally, the availability of two data points for the same household also enables validation of this data across specific variables, although in the absence of the "ground truth", the research team had to make a subjective judgement based on their own contextual understanding on which response can be considered the "true" value. Based on lessons learned from this approach, this methodology is also currently being adapted to be used and further tested for similar household surveys to be rolled out in Afghanistan and Lebanon in August 2023.

- **Validation of key informant interview data – Afghanistan, DRC and South Sudan, 2022:** Since 2013, IMPACT has used a key informant (KI) based data collection methodology, internally referred to as 'Humanitarian Situation Monitoring' (HSM) to collect quantitative data for areas where standard household survey approaches cannot be implemented on a regular basis. In order to validate the reliability of the data produced with this method, while also testing some assumptions underlying its research design, a validation study was launched by IMPACT at the end of 2021. The key research question for this study was to understand to what extent HSM data is truly indicative of humanitarian needs in assessed areas. Since the 'ground truth' is not really known or possible to establish for these areas, data from household-level MSNAs conducted by IMPACT in the same areas was assumed to be the superior comparator for the purpose of this study. Although the final analysis is still ongoing and publication forthcoming in August 2023, preliminary findings indicate that HSM data is able to provide some accurate indication of humanitarian needs in assessed areas, especially for specific food security, protection, and water, sanitation and hygiene (WASH) indicators in South Sudan. On the contrary, findings are less conclusive for Afghanistan, indicating the need for further research at a more granular level across multiple contexts. A similar approach was also previously tested by IMPACT to determine reliability of KI data in specific urban areas of Jordan, Niger and Uganda in 2017 and 2018.[18]

Although validation studies like these can be useful to help determine the extent of measurement errors, **there are certain challenges and limitations to their implementation, especially for humanitarian contexts** where researchers constantly face challenges with time, access and limited availability of up-to-date, reliable secondary data. These challenges and limitations are summarized below:

- **The first limitation of this approach is the methodology itself**, since availability of comprehensive and reliable validation data is an exception rather than a rule (Schennach, 2016). In other words, unless we know the "true" values of the variables we want to measure and verify, or at minimum have some way of obtaining a more superior comparator, such external validation is not really possible.
- **Secondly, the approach assumes, often incorrectly, that the validation data has higher validity than survey data and is completely without error**, and thus could over-estimates the level of response error within the survey (Bound et al., 2001). This is not necessarily true, and presumably error-free sources such as administrative data or payroll records can also include errors. Even if the validation data is truly

[17] Publication available upon request.

[18] For an overview of this approach and preliminary findings from Uganda, please refer to Annex 1 (page 52) of this publication: IMPACT Initiatives (2018), *Area-based Assessments with Key Informants: A Practical Guide*

error-free, these studies are not always helpful as the extent of the error is not summarized in a way that can guide the analysts' final conclusions from the data (Bound et al., 2001).

- **Thirdly, validation studies rely on the presumption that the survey measures exactly the same construct as the validation data**, which also may not always be the case (Bound et al., 2001). Moreover, the usefulness of validation studies can also vary based on the variable and type of topic being studies. For example, in their research on labour economic survey data, Duncan & Hill (1985) found that errors were more likely for less salient features such as unemployment, working hours and sick time, while being less likely for the more salient aspects of current employment situation, such as pension plans and health insurance schemes. They also found that the extent of the error could vary depending on the recall period.
- **Finally, conclusions from validation studies are not necessarily scale-able, since the data collected in one context may not apply to another**, and existing methodological research has suggested that for many variables, including reported earnings and employment data, the extent of measurement error can be quite context-dependent (Bound et al., 2001). This was also found to be the case for IMPACT's ongoing validation study for key informant data; as mentioned above, preliminary findings from Afghanistan are less conclusive than South Sudan, indicating the need for further research across a wider range of contexts.

When external validation is not possible due to the challenges outlined above, researchers still can and should ensure some level of triangulation is included during the data processing analysis, either by using a variety of techniques to record observations or using multiple data sources to verify conclusions (e.g. mixed methods research or joint analysis exercises). Further confirming this, when representatives from IMPACT's research and data teams across different contexts were asked what they have found to be the most effective way to minimize systematic errors across their respective contexts, the most commonly reported response, especially among respondents from the MENA region, was "*triangulation i.e. using multiple data sources to verify observations*". Two respondents also perceived the use of additional qualitative data sources to be an effective measure to validity of reported survey data, including interviews with community leaders and subject matter experts where feasible.

Based on their experiences conducting public health research across different humanitarian contexts, Karadag et al. (2021) also highlight the importance of "*collaboration with local researchers and institutions and acknowledging local knowledge and experience are vital for better public health research and practice outcomes*". Similarly, Guha-Sapir & Scales (2020) note that when "*large-scale displacement, destruction of roads and access channels, and whole-family deaths… complicate population sampling methods",* the use of facility data for public health research becomes *"a defensible choice*". Till date, they have reportedly undertaken four studies using facility-based data complemented by qualitative techniques, and even though this can have its limitations, the strengths of using this data were perceived to outweigh the potential selection bias, especially after initial preparatory work clearly indicated that population-based sampling presented too many obstacles. Additionally, since self-reported morbidity can have reliability concerns due to response (and recall) biases, use of facility-based data, especially patient records, can help researchers to ascertain objective diagnosis of injuries and cause of death from clinical sources, thus improving data quality and strengthening the estimation of mortality and morbidity due to disasters (Guha-Sapir & Scales, 2020).

Another important form of triangulation to help overcome some of the barriers faced in humanitarian data collection contexts is the

use of participatory research approaches and the involvement of local communities and crisis-affected populations throughout the research process. There is sufficient literature showing that "*community participatory research approaches that include working with community leaders, cultural mediators, and civil society organizations*", combined with using peer-to-peer reviewed methodologies in data collection and analysis, can greatly increase the quality of research with crisis-affected populations, especially migrants and refugees (Karadag et al., 2021). Indeed, upon analysing lessons learned from over 20 different research exercises conducted across diverse humanitarian crises, populations and topics of study, Mistry et al. (2021) found that a common and important element of almost every study was the engagement of affected populations and local communities in the research process. On the other hand, after conducting a case study of three different mortality and nutritional survey reports published in North Kivu, Democratic Republic of Congo between January 2006 and January 2009, Grais et al. (2009) found that although reporting against minimum criteria was generally good, all surveys failed to consider contextual factors important for data interpretation and improved data quality. These contextual factors can only be properly accounted for if research approaches are participatory and inclusive of local communities throughout the process.

For instance, based on their experiences of conducting research with refugees in urban contexts, Karadag et al. (2021) concluded that the quality of their data collection and analysis was greatly enhanced because of having included refugees as part of the data collection teams. Similarly, in 2019, IMPACT's team working on the Rohingya refugee response in Bangladesh's Cox's Bazar district conducted a pilot study with the participation of Rohingya enumerators in data collection activities; the study found that both the community reception of the data collection exercise, as well as consistency of responses and inputs to the survey, did vary based on whether respondents were speaking to Rohingya or Bangladeshi data collectors.[19]

Even if participation throughout the research process is not always feasible due to limited time and access, ensuring local community members are consulted at least during the data processing and analysis stage is important to try and ensure survey data collected truly reflects the situation on the ground. For instance, in August 2019, IMPACT's team in Somalia, in partnership with UN-OCHA and the Africa Voices Foundation (AVF), delivered an innovative intervention which deployed AVF's Common Social Accountability Platform to disseminate findings from the household-level MSNA survey to local communities via interactive radio programmes. The goal of this was to gather feedback on the survey findings from crisis-affected populations, while also sparking a broader public dialogue on the priorities of the humanitarian response in Somalia. In a two-week rapid consultation, AVF heard from more than 8,000 people who engaged directly with the programme, expressing their feedback in their own words, thus helping not only to verify the survey findings but also to directly inform the design of the humanitarian response.[20]

Similarly, in Iraq, the International Organisation for Migration (IOM) conducted an exploratory study to pilot "*a novel form of collective intelligence that enables returnees in Iraq to validate and improve processes for the collection and analysis of data related to the conditions in their local area*" (Trigwell et al., 2022). The study used digital data collection channels, including online surveys, to engage a more diverse cross-section of returnees and validate findings from key informant data collection. Specifically, location-specific conclusions from previous data collection

[19] See also: IMPACT Initiatives (2019), *Participation of Rohingya Enumerators in Data Collection Activities: Findings from a Pilot Assessment in Cox's Bazar, Bangladesh.* https://rb.gy/vb121

[20] See also: UN-OCHA et al. (2019), *Amplifying community voices in humanitarian action in Somalia.* https://rb.gy/kygpu

activities were shared back with respondents to confirm or reject, and respondents were asked to provide open-ended qualitative inputs to further explain their response. While the limited number of respondents in each location prevents the attainment of any statistically significant findings, the approach still enabled more meaningful participation of returnees and helped bring in contextual knowledge necessary to not only validate data, but also to improve data collection tools for the future (Trigwell et al., 2022).

4.3 Other measures being explored by IMPACT to address measurement errors in humanitarian settings

In addition to the different measures that have been discussed so far, **IMPACT is currently also exploring specific measures to address two commonly encountered errors from its own experiences across different humanitarian contexts:** 1) systematic response errors and potential data falsification due to enumerator fatigue and / or demotivation; and 2) random errors due to cluster sampling design. This sub-section will discuss both of these in more detail.

- **Addressing systematic response errors and potential data falsification due to enumerator fatigue / demotivation**

As previously noted in section 3 above, a key source of systematic response error can be fatigue or lack of motivation on the side of the enumerator, which not only leads to unintentional data entry errors but also intentional falsification of survey responses. The risk of this is especially high in humanitarian contexts where access limitations makes direct oversight of data collection especially challenging, and researchers often have to remotely train and supervise data collection teams from afar. It is therefore possible that entire surveys do not reflect the views of an interviewed household member, but are answered randomly by the enumerators. However, identifying such falsified data is challenging, and sometimes, data points have to be deleted *a posteriori* to follow a conservative approach to the treatment of falsified data points i.e. deleting all records from a specific enumerator showing suspicious trends, thus losing large chunks of potentially correct information.

Because of this, **systematically monitoring enumerator behaviour is a minimum standard required from the data cleaning process for every survey implemented within IMPACT**, and each team is required to document the outcomes of this within the final cleaning log of their dataset (see example in Figure 1 of the Appendix). Additionally, for specific food security outcome indicators, IMPACT also uses an internal R script to check distribution of responses per enumerator for each indicator (see example visualisation in Figure 2 of the Appendix). In 2020, IMPACT's team in Syria also explored the use of silhouette analysis to monitor enumerator survey behaviour and detect potentially falsified surveys.[21] By assuming the dataset is clustered by enumerator IDs, the silhouette value was calculated using the Gower distance between surveys.[22] Since a silhouette value close to 1 indicates that the entries of the cluster are very similar to each other and dissimilar from entries of other clusters, a 'flag' was raised if the silhouette value was above 0.5 for any of the clusters / enumerators.

In addition to the above, **in 2020 and 2022, IMPACT collaborated with a team of students from the Swiss Federal Institute of Technology (ETH) in Zurich,** through the annual Hack4Good programme, **to explore**

[21] Silhouette analysis is a method of interpreting and validating consistency within clusters of data. In other words, a silhouette value measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

[22] The Gower distance is essentially a measure of data similarity, and can be used to calculate how different or similar two records are. See also: Anand, D. (2020). Gower's Distance. *Medium blog post.* https://rb.gy/ebp4y

more efficient, automated solutions for the systematic detection and treatment of falsified survey data. In 2020, using IMPACT's annual MSNA dataset from Afghanistan which comprised of approximately 40,000 household survey records, and treating the team's cleaning and deletion logs as the "ground truth", the students developed a supervised machine learning algorithm (Extreme Gradient Boosting) to detect falsified data points and / or surveys. The algorithm showed some promising results for predicting data entries which had been cleaned due to potential falsification, but since it was developed specifically for the dataset from Afghanistan it was not scale-able and could not be tested easily for other contexts.

Building on this initial exploration, in 2022, a new team of students were tasked to look at IMPACT's MSNA dataset from Burkina Faso with a similar objective. However, this time, to ensure that the outputs are easily adaptable to other contexts, the team focused on the KoBo audit files to detect anomalies in enumerator behaviour,[23, 24] rather than the dataset itself, since the format of audit files is similar for all data collection using the KoBo platform.[25]

By using the audit files to analyse how each enumerator was interacting with the data collection app during the survey, the team was able to identify a list of features that could be the sign of anomalies in enumerator behaviour (Johansson et al., 2022).[26] Examples of these features include 'duration' (i.e. the duration of time the enumerator spent on the survey) and 'resume_count' (i.e. number of the times the survey was paused and then resumed by the enumerator). Based on these features, an Isolation Forest algorithm was used to identify anomalous surveys.[27]

Following this approach, the team developed a series of 23 features to characterize possible way of detecting anomalies while enumerators are filling out the questionnaire. For an example of these features and their descriptions, see Table 1 in the Appendix. The tool produced is essentially able to use the raw dataset and audit files to automatically calculate all features and outputs for the anomalous survey responses. A code further plots comparisons of the feature distributions for survey responses to identify the suspicious ones. Finally, 1) a table is produced showing a score per survey (lower the score, the more suspicious the survey) and the primary reason why the survey is flagged (see Table 2 in the Appendix for an example); and 2) a graph is also produced showing the percentage of anomalies by enumerator and region (see Figure 3 in the Appendix for an example).

Although the tool is promising, it comes with certain challenges and limitations (Johansson et al., 2022). Firstly, since it is almost

[23] Read more about this platform here: https://www.kobotoolbox.org/

[24] All of IMPACT's survey data collection is currently done with the KoBoToolbox data collection tool (read more here). A few years ago, KoBoToolbox introduced a functionality called 'audit' that records the enumerator behaviour as he/ she goes through each survey. Specifically, it provides insight on how enumerators were interacting with the data collection app when filling out their survey responses. This can then be downloaded as an Excel CSV file which provides detailed information per survey, including every action taken and the time spent on starting survey, going to next question, leaving survey, jumping between questions, etc. Read more about this audit function here

[25] For each survey, there are at least 4 columns in the audit files: 1) event – what type of action is done on the form e.g. start, viewing question, jumping question, exiting the form, saving the form, location tracking, etc.; 2) node – what is the question for this event; 3) start – the time the event started; and 4) end – the time the event ended.

[26] For finding such features, the team performed a three-step brainstorming process, answering the following questions: 1) What are the ways in which an enumerator can try to falsify a survey response?; 2) What patterns will that generate in the time information?; and 3) What features can capture these patterns?

[27] At each step, the algorithm selects a feature and a random value at which to "cut" or separate that data. It continues making random cuts until the data points are isolated from one another. The average number of cuts it takes to isolate a given data point is then a sign of how anomalous or typical that data point is. Anomalous points require, on average, fewer cuts to be isolated than normal data points.

impossible to be certain which survey or response is actually falsified, using supervised learning techniques to classify fraudulent responses is quite challenging. For the same reason, it also proved difficult to objectively assess the performance of the methodology. Another difficulty arises from the size of the dataset itself; in order for the tool to function properly, it requires a large amount of data which means it might not work for smaller data collection exercises or too early in the data collection process. Finally, substantial familiarity with the survey questionnaire and the data collection context is needed to properly interpret and follow-up on the results, especially since every flagged survey is not necessarily incorrect. Because of these limitations, further exploration and testing is needed to improve the overall performance of this tool and its ability to accurately detect falsified data points or surveys.

Despite these limitations, the tool has promising potential for more efficient and systematic detection of falsified data. IMPACT is currently looking at how this tool can be adapted and developed further, especially since the consistent format of audit files for all data collection projects makes it easily transferable across contexts. One potential use case could be for early in the data collection process, to detect anomalies in enumerator behaviour and adapt plans accordingly (e.g. retraining or re-recruiting teams as appropriate). As a first step in the next months, IMPACT will look at testing the tool on another dataset from a similar context (e.g. Mali), specifically to review the list of features. Once it has been tested across different contexts, the solution, if applicable, will be integrated within IMPACT's global Data Cleaning Guidelines.

- **Addressing random errors caused by Cluster Sampling design**

As previously discussed in section 3, a key source of random errors is errors caused by sampling, and this is especially true in humanitarian contexts where inability to build proper sampling frames and implement standard randomisation techniques makes such errors even more likely. Within IMPACT, a large proportion of surveys, especially those implemented at larger crisis-wide levels, use a Cluster Sampling methodology, both to overcome difficulties linked to lack of comprehensive sampling frame and household lists, as well as to ease logistical planning with limited time and access available. For the design of such surveys, an estimated design effect[28] is used based on a standard intra-cluster correlation (ICC) of 0.06, which was calculated based on past surveys from a few years ago.[29] However, this does not account for any variance in design effect by variable, population group and geographical area. As a result, the sampling design is currently almost blind to random errors, thus potentially lowering the precision of findings produced.

To take the opportunity of increasing standardisation of survey tools and methods (especially for the MSNA), **in 2022, IMPACT collaborated with Statisticians without Borders (SWB) to review best practices for estimating design effect and ICC for household surveys, and provide recommendations for improving Cluster Sampling design within IMPACT.** A team of SWB volunteers looked at IMPACT's MSNA datasets from 2022 across five contexts -

[28] In survey methodology, this is a measure of the expected impact of a sampling design on the variance of an estimator for some parameter. Within IMPACT, the formula used to calculate design effect is $neff = n\,(\,1 + (\,M - 1\,)\,ICC\,)$; where $neff$ = effective sample size, n = unadjusted sample size, M = average sample size per cluster, ICC = intra-cluster correlation

[29] The ICC is essentially a descriptive statistic which describes how strongly units in the same group resemble each other. It is essentially a type of correlation, and commonly used to quantify the degree to which individuals with a fixed degree of relatedness (e.g. in the same village) resemble each other in terms of a quantitative trait (e.g. access to water points). A higher ICC indicates more similarity of households within the cluster; as such, high ICC implies high design effect and thus lower precision of findings.

Burkina Faso, Central African Republic (CAR), Lebanon, Occupied Palestinian Territories (OPT) and South Sudan - and calculated ICCs for 29 standardised variables available within each of these datasets. Both weighted and unweighted ICCs were calculated wherever possible, and although no geographical disaggregation was done, separate ICCs were also calculated for host community, internally displaced and returnee population groups, where applicable.

A couple of interesting results were observed based on this (SWB, 2023). Firstly, across population groups, most contexts had similar ICCs with a few exceptions. However, ICCs were found to vary quite a lot between variables (see Table 3 in Appendix for an example from OPT). In OPT for example, ICC was found to vary between 0.003 for education indicators like children dropping out of school and as high as 0.84 for WASH indicators like primary source of drinking water for the household. Similarly, in CAR, ICCs ranged from 0.5 for the measuring households' access to healthcare, and 0.93 for type of shelter occupied by the household. In addition to confirming that design effect does vary by variable, these findings also show that certain variables (and the underlying experiences they are trying to measure), including access to water and type of shelter, are spatially correlated and might not necessarily need to be captured through household surveys.

Although these results are interesting, they are indicative only at this stage and IMPACT is not able to take concrete action on them just yet, especially due to limited documentation available for each sampling approach. However, as an immediate next step, IMPACT will look into implementing the ICC calculation for all standardised variables within the MSNAs conducted in 2023, to continue this line of analysis. Additionally, these findings will also inform an internal review on the efficiencies of current sampling approaches, specifically in terms of the use of household survey method as opposed to other methods (e.g. spatial analysis or remote sensing) for collection data on spatially correlated variables.

5 Conclusion

Over the past few years, survey research has emerged as a powerful method to collect and analyse data on crisis-affected populations in order to inform more evidence-based planning and delivery of humanitarian aid. However, designing and implementing robust survey methodologies in humanitarian settings can be especially challenging, and while well-established examples do exist, **the complex operational contexts associated with humanitarian settings introduces unique scientific challenges and conditions** that distinguish them from standard research practices. There are three key challenges specific to survey research in humanitarian contexts that are worth noting: 1) limited time and accessibility, 2) additional ethical considerations of researching vulnerable, and often traumatized, crisis-affected populations, and 3) lack of reliable, up-to-date secondary data sources, including standardized administrative methods for record keeping, data sharing and dissemination.

As a result of these challenges, surveys implemented in humanitarian contexts are especially prone to measurement errors, both random and systematic, which need to be properly understood and addressed to ensure the timely and consistent production of high-quality data for humanitarian decision-making. Firstly, since implementation of probability sampling can be especially challenging in humanitarian contexts due to lack of time, access and prior information on the population of interest, survey research in such contexts is prone to random error affecting the overall precision of data produced. Secondly, survey research in such contexts are also equally, if not more, prone to systematic errors, especially response errors arising from a range of sources including problematic questionnaire design, social desirability biases, enumerator fatigue and/ or

de-motivation, as well as specific conditions in the data collection environment. These types of errors can often be more complex, varied and difficult to detect, with serious effects on the overall accuracy of data that is produced.

As such, **quite a lot of effort is required to systematically detect the types and sources of measurement errors affecting survey data in humanitarian contexts, as well as to ensure these are effectively addressed as much as possible**. First and foremost, regardless of the type of error, the first step in ensuring quality of survey data is to identify the sources of errors in a given data collection context. In order to do this, a robust data cleaning process needs to be set up and followed for each survey. This is especially important for the complex operational data collection environments that are characteristic of humanitarian contexts, where as a result of access and security barriers, researchers often have to supervise data collection remotely and have limited control over the implementation of the survey on the field. Secondly, while treatment of random errors is perceived to be relatively more straightforward because it primarily requires increasing sample sizes and ensuring implementation of robust randomised sampling techniques, this is easier said than done in humanitarian contexts where accurate population data to design sampling frames is rarely available, and sampling errors are unavoidable due to access and security restrictions. As such, in these contexts, more creative solutions need to be considered as appropriate, including the use of tailored GIS-based sampling solutions, as well as use of mixed methods research with qualitative approaches. Finally, treatment of systematic errors, especially response errors, tends to be the most challenging due to the varied and complex nature of their sources, but examples and best practices are available that can be considered for humanitarian contexts. For instance, the use of triangulation measures can be quite effective to determine validity of collected survey data, both in terms of structured validation study designs as well as more inclusive and participatory research designs to ensure meaningful engagement of crisis-affected communities.

Overall, while in an ideal scenario consistent efforts need to be made to reduce random and systematic errors simultaneously across all surveys, researchers will often need to make a trade-off between selecting large samples to minimize random sampling error, or **focusing the limited time and resources available on a smaller sample to ensure better controls on the data collection process**, higher response rates and more accurate responses (Assael & Keon, 1982). Given the operational complexities surrounding survey research in humanitarian contexts, and limited time, resources and access available for the implementation of robust sampling techniques, IMPACT has till-date focused more on the latter as explained through the examples presented in this paper.

Finally, even if efforts are made to minimize error sources during data collection, researchers are seldom able to measure total survey error during the analysis stage due to limited availability of valid data and confirmatory information sources for all subjects (Assael & Keon, 1982). This further complicates external validation when the "ground truth" is assumed rather than known, as demonstrated through the two examples IMPACT is currently exploring for validation of KI data, and development of automated solutions for detection of falsified survey data. Nonetheless, in order to ensure the continued production of high-quality survey for effective and evidence-based humanitarian decision-making, survey researchers operating in such contexts need to continue exploring effective measures and techniques to address commonly known measurement errors and improve the overall accuracy and precision of survey data produced. Recent advances in data science technologies can also be leveraged for this purpose, as demonstrated through IMPACT's ongoing exploration to develop supervised machine learning solutions for better and more efficient detection of falsified data through all its surveys.

In sum, as humanitarian crises increase in scale and severity across the globe, the need for accurate and reliable data to inform more evidence-based humanitarian action has become more important than ever. The present reality is that humanitarian needs around the world are higher than ever before, but humanitarian funding is unable to keep up with this growing scale and severity of needs. In this difficult global environment, taking evidence-based decisions around planning and prioritization is therefore becoming increasingly necessary. At the same time, there is a unique opportunity right now with the humanitarian community undergoing a so-called "data revolution", as aid organisations are collecting, producing and sharing more data than ever before. Collectively understanding the opportunities this brings, while addressing the limits of existing data production processes, is therefore key to ensuring that data-driven aid action continues to be based on the most relevant and best possible quality of information available.

References

1. Alwin, D. F. (1989). Problems in the estimation and interpretation of the reliability of survey data. *Quality & Quantity*, 23(3–4). https://doi.org/10.1007/bf00172447
2. Assael, H., & Keon, J. P. (1982). Nonsampling vs. Sampling Errors in Survey Research. *Journal of Marketing*, 46(2), 114–123. https://doi.org/10.1177/002224298204600212
3. Bhandari, P. (2021). Random vs. Systematic Error: Definition and Examples. *Scribbr blog post.* https://rb.gy/wl6k5
4. Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, 74(5), 817–848. https://doi.org/10.1093/poq/nfq058
5. Bound, J., Brown, C. R., & Mathiowetz, N. A. (2001). Measurement Error in Survey Data. *Handbook of Econometrics* (pp. 3705–3843). Elsevier BV. https://doi.org/10.1016/s1573-4412(01)05012-7
6. Das, N. (2023). Putting needs first: Making effective use of data and analysis for a needs-based approach to humanitarian action. *2023 UN World Data Forum blog post.* https://rb.gy/trs23
7. Duncan, G. J., & Hill, D. (1985). An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data. *Journal of Labor Economics*, 3(4), 508–532. https://doi.org/10.1086/298067
8. Fogarty International Center (2021). *Scientists describe challenges and lessons learned when conducting research in humanitarian crises.* https://rb.gy/57fle
9. Fowler, F. J. (2009). *Survey Research Methods*. SAGE
10. Grais, R. F., Luquero, F. J., Grellety, E., Pham, H., Coghlan, B., & Salignon, P. (2009). Learning lessons from field surveys in humanitarian contexts: a case study of field surveys conducted in North Kivu, DRC 2006-2008. *Conflict and Health*, 3(1). https://doi.org/10.1186/1752-1505-3-8
11. Guha-Sapir, D., & Scales, S. E. (2020). Challenges in public health and epidemiology research in humanitarian settings: experiences from the field. *BMC Public Health*, 20(1). https://doi.org/10.1186/s12889-020-09851-7
12. IMPACT Initiatives (2019-2022). *Lessons Learned Documentation from REACH Multi Sector Needs Assessments*.
13. IMPACT Initiatives (2020). *Data Cleaning Guidelines for Structured Data.* https://rb.gy/61qno
14. Inter-agency Standing Committee (2021). *The Grand Bargain 2.0: Endorsed framework and annexes*. https://rb.gy/0nwkz
15. Johansson A.T., Rothschuld, J., & Zehr, A. (2022). *Hack4Good Fourth Edition Final Report – Making an IMPACT*. ETH Zurich Hack4Good Initiative in collaboration with IMPACT Initiatives.
16. Kapteyn, A., & Ypma, J. Y. (2007). Measurement Error and Misclassification: A Comparison of Survey and Administrative Data. *Journal of Labor Economics*, 25(3), 513–551. https://doi.org/10.1086/513298
17. Karadag, O. O., Kılıç, C., Kaya, E., & Üner, S. (2021). Challenges and lessons learned in mental health research among refugees: a community-based study in Turkey. *BMC Public Health*, 21(1). https://doi.org/10.1186/s12889-021-11571-5
18. Khan, S. M., & Hancioglu, A. (2019). Multiple Indicator Cluster Surveys: Delivering Robust Data on Children and Women across the Globe. *Studies in Family Planning*, 50(3), 279–286. https://doi.org/10.1111/sifp.12103
19. Meyer, B. D., Mok, W. K. C., & Sullivan, J. (2015). Household Surveys in Crisis. *Journal of Economic Perspectives*, 29(4), 199–226. https://doi.org/10.1257/jep.29.4.199
20. Mistry, A. S., Kohrt, B. A., Beecroft, B., Anand, N., & Nuwayhid, I. (2021). Introduction to collection: confronting the challenges of health research in humanitarian crises. *Conflict and Health*, 15(1). https://doi.org/10.1186/s13031-021-00371-8
21. Saris, W. E., & Revilla, M. (2015). Correction for Measurement Errors in Survey Research: Necessary and Possible. *Social Indicators Research*, 127(3), 1005–1020. https://doi.org/10.1007/s11205-015-1002-x

22. Schennach, S. M. (2016). Recent Advances in the Measurement Error Literature. *Annual Review of Economics*, 8(1), 341–377. https://doi.org/10.1146/annurev-economics-080315-015058
23. Statisticians without Borders (2023). *Survey Sampling Methodology - Project Update Presentation*. In collaboration with IMPACT Initiatives.
24. Trigwell, R., J. Phillippo-Holmes, E. Zambrano, J. Bahn, P. Hirani and E. Griesmer (2022). *Validating Humanitarian Data Analysis Through Collective Intelligence: A Pilot Study.* International Organization for Migration, Geneva. https://rb.gy/3hele
25. UN-OCHA (2022). *Global Humanitarian Overview for 2023*. https://rb.gy/1avcp

Appendix

Figure 1: *Example summary of enumerator performance analysis to be included within cleaning log for each dataset within IMPACT*

Dataset

Number of surveys collected by enumerators

Enumerator ID	Number
enu1	23
enu10	22
enu2	27
enu3	31
enu4	34
enu5	35
enu6	32
enu7	30
enu8	26
enu9	24
(blank)	
Grand Total	**284**

Cleaning log

Number of changes by enumerators

Enumerator ID	Number
enu1	2
enu2	2
enu4	2
enu6	2
enu7	1
enu9	1
(blank)	
Grand Total	**10**

Number of changes by enumerators filtered by issues

Type of Issue (Select from dropdown list): (All)

Enumerator ID	Number
enu1	2
enu2	2
enu4	2
enu6	2
enu7	1
enu9	1
(blank)	
Grand Total	**10**

Deletion log

Number of deletions by enumerators

Enumerator ID	Number
enu1	4
enu10	4
enu2	4
enu3	4
enu4	4
enu5	3
enu6	5
enu7	2
enu8	3
enu9	1
(blank)	
Grand Total	**34**

Number of deletions due to time by enumerator

Issue: (All)

Enumerator ID	Number
enu1	4
enu10	4
enu2	4
enu3	4
enu4	4
enu5	3
enu6	5
enu7	2
enu8	3
enu9	1
(blank)	
Grand Total	**34**

Figure 2: *Example visualization output from R script looking at distribution of responses by enumerator for the Reduced Coping Strategy Index (rCSI) indicator*

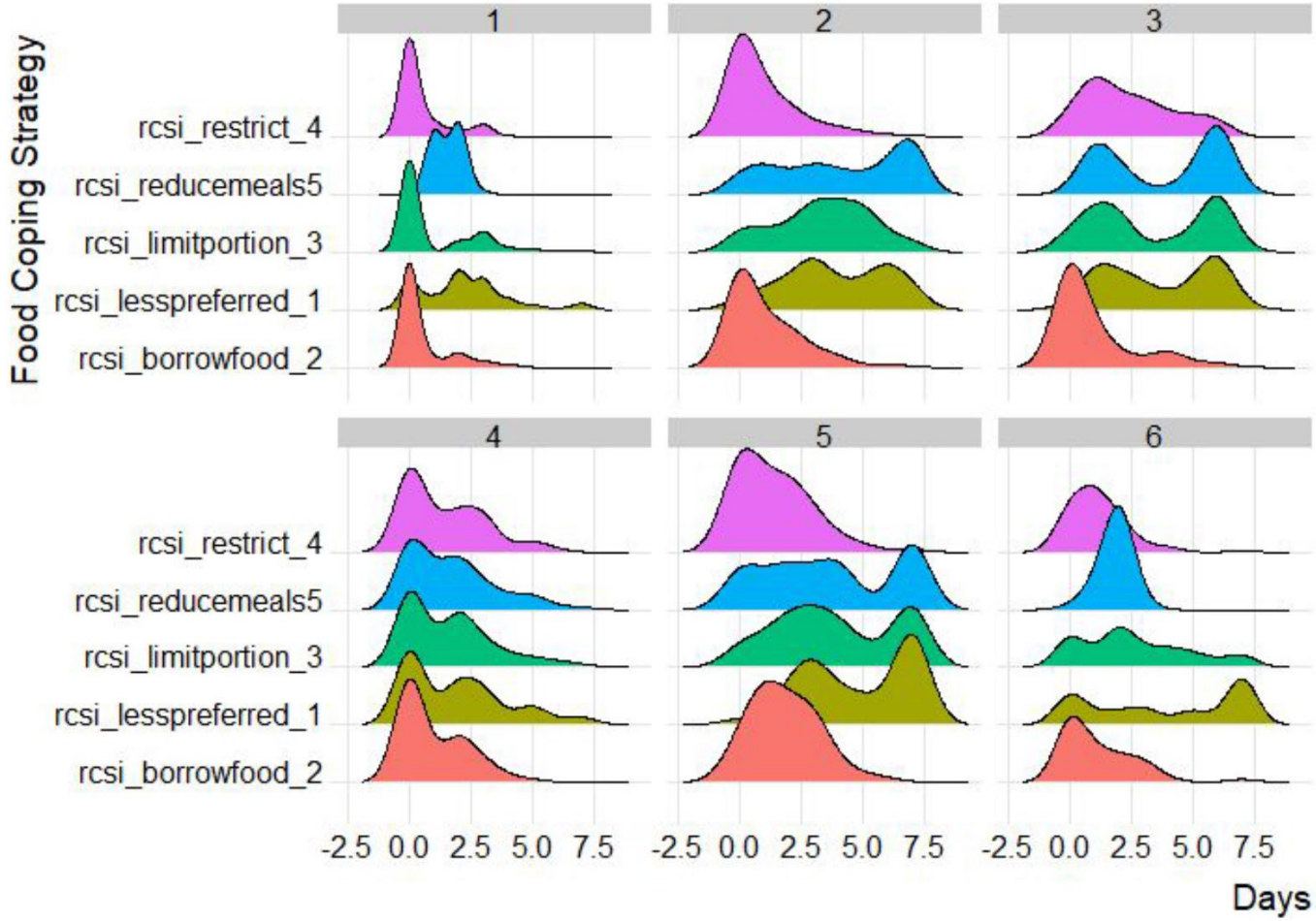

Table 1: *Example of features produced by applying Isolation Forest Algorithm to enumerator behaviour data within KoBo audit files (Johansson et al., 2022)*

Feature Name	**Technical Description**	**Intended Goal**
duration	This calculates the duration of time that the survey was worked on (in minutes). Any long pauses where the survey was paused and later resumed are not counted towards this time.	Look for surveys that may have been too hastily completed, indicating that the questions were not actually asked.
resume_count	Number of times the survey was paused and then later resumed, as evidenced by a "form resume" node in the audit file.	This helps distinguish strange behavior.
constraint_errors	Number of times the "constraint error" event appears in the audit file.	Similar logic to the prior feature.
median_seconds , deindexed_median_seconds	This computes the median time in seconds for a specific node, over all surveys.	This helps compute e.g. relative pace.
enum_active_se c_on_surveys_that_day	This computes the number of seconds spent by the enumerator on surveys the day the survey was started.	This helps identify when significantly shorter time is spent on surveys, which could indicate an attempt to rush through the minimum number of required surveys in order to get paid.
active_fraction_b efore_long_break	This computes the fraction of active seconds spent on this survey that occurred before its first long break.	This helps keep indicate long breaks followed by much of the survey being filled out quickly.

Table 2: *Example of table produced by machine learning tool to detect surveys with suspicious enumerator behaviour as per KoBo audit files (Johansson et al., 2022)*

A	Q	R	S	T
uuid	global_enum_id	score IF	first_important_feature	second_important_feature
3a46dee8-00c3-47c8-bc14-e70019e90d38	dori_105	-0.76	variance	duration
c650135d-a4cf-4645-801e-50e7367a1960	fada_401	-0.73	constraint_backtracks	largest_relative_pace_increase
3d7c6650-689d-4c99-bca5-21af3323d3f9	ouahigouya_303	-0.71	resume_count	constraint_errors
f00c674d-24ba-4616-bd74-296e40bb505d	fada_415	-0.71	variance	enum_active_sec_on_surveys_
d7f7b03d-0bae-4eca-9776-9b5e09dc2b04	fada_415	-0.68	enum_active_sec_on_surveys_t	duration
882cee33-a645-4ce0-a47f-9e37d2170c3f	fada_415	-0.68	constraint_backtracks	active_fraction_before_long_br
079a016e-33ca-4517-8e96-a4f7331dca71	dedougou_513	-0.67	constraint_errors	resume_count
1903bc50-b49d-4989-99f8-83a965b1f9e3	dedougou_511	-0.67	resume_count	duration

Figure 3: *Example of graph produced by machine learning tool to visualize proportion of suspicious surveys per enumerator and region (Johansson et al., 2022)*

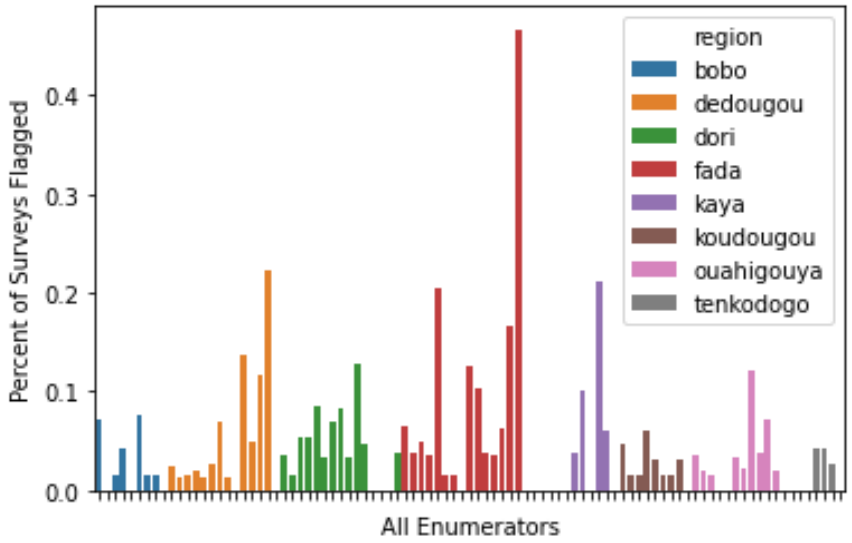

Table 3: *Differences in estimated ICC for some variables of interest in the Occupied Palestinian Territories (Statisticians without Borders, 2023)*

Variable	$\hat{\delta}$
Food consumption score category acceptable	0.094
Household hunger scale	−0.077
Needs healthcare	0.232
At least one household member disabled	0.134
Respondent age	−0.142
Respondent gender female	−0.226
Size of the household	−0.029
Head of household female	−0.073
Number of men in the household	−0.101
Number of women in the household	0.074
Number of boys in the household	−0.094
Number of girls in the household	−0.093
School age children	−0.068
Number of adults	0.017
Number of children	−0.092
At least one of the members of the hh reporting difficulty seeing	0.114
At least one of the members of the hh reporting difficulty hearing	0.061
At least one of the members of the hh reporting difficulty walking	0.128
At least one of the members of the hh reporting difficulty remembering	−0.045
At least one of the members of the hh reporting difficulty selfcare	0.243
At least one of the members of the hh reporting difficulty communication	−0.094