Statistical Commission
Fifty-fourth session
28 February – 3 March 2023
Item 3(c) of the provisional agenda
Items for discussion and decision: Household Surveys

Background document
Available in English only

Guidance Note on Assessing and Minimizing the Impact of a Crisis on Survey Quality: Approaches Learned from the COVID-19 Pandemic

Prepared by

Inter-Secretariat Working Group on Household Surveys
United Nations Statistics Division
ECLAC Statistics Division

Table of Content

List of tables

List of figures

Acknowledgement

1. Initial draft of the Guidance Note was prepared by David Marker, consultant to the United Nations Statistics Division. The drafting process was overseen by Haoyi Chen, Coordinator for the Inter-Secretariat Working Group on Household Surveys, who was also the lead author for the Guidance Note. The Guidance Note was jointly authored by Andres Gutierrez Rojas, Regional Advisor of the United Nations Economic Commission for Latin America and the Caribbean. The authors also benefited greatly from Kieran Walsh from International Labour Organization (ILO) for his guidance and suggestions throughout the drafting process. Special gratitude goes to Charlotte Taglioni and Nemi Okujagu for their support in the drafting process.

2. The Guidance Note was produced under the direction of Gero Carletto, Manager of the Data Production and Methods under the Development Data Group at the World Bank and co-Chair of the Inter-Secretariat Working Group on Household Surveys and Francesca Perucci, Assistant Director of the United Nations Statistics Division.

3. This Technical Guidance Note is the result of collective efforts, involving a wide range of contributors with extensive experience in household survey operations. Sincere appreciation goes to the following experts who reviewed and/or provided technical advice: Kieran Walsh, International Labour Organization (ILO); Salilkumar Mukhopadhay, Central Statistics Office, India; Seoyoung Kim, KOSTAT; Danielle Groffen, Statistics Netherlands; Suffiea Ibrahim and Rania Abu Ghaboush, Palestinian Central Bureau of Statistics; Afsaneh Yazdani, United Nations Economic and Social Commission for Asia and the Pacific; Jessamyn Encarnacion, UN Women; and Kelly L'Engle, University of San Francisco.

Executive summary

4. The COVID-19 pandemic has turned everyone's life upside down, including the world of official statistics. Household surveys, arguably, are among the most impacted areas in official statistics, given that a large proportion of countries relied heavily on face-to-face interviewing before the pandemic. The pandemic, however, has also become a catalyst for innovation. In addition to the use of nontraditional sources as indicated by countries in rounds of surveys carried out by the UN Statistics Division and the World Bank (United Nations Statistics Division and World Bank, 2020a and 2020b), national statistical offices also adapted quickly in changing data collection mode from face-to-face to remote, either through telephone or web interviewing, to meet the urgent data demand. As compiled by the Inter-Secretariat Working Group on Household Surveys, almost all countries have had at least one remote data collection during the pandemic, either independently or with the support of ISWGHS members (ISWGHS, 2022a).

5. Under normal circumstances, developing a new survey instrument or introducing a new survey element takes months, if not years, of testing and experimentation. As countries thriftily adopted the remote data collection during the pandemic, national statistical offices raised two questions in a global survey carried out by the Inter-Secretariat Working Group on Household Surveys with its stakeholders. First, how have these rapid adaptations in survey operations impacted survey data quality? And second, what have we learned from the experiences during the pandemic, and how can we be better prepared for the next crisis?

6. The Guidance Note aims to respond to these two questions. It documents many changes that were introduced to survey operations during the pandemic, including the mode of data collection and as a result, the availability of relevant sampling frames; questionnaire design and other changes such as training, supervision, quality assurance process, and data collection and processing protocols. Methods to assess and reduce the impact of COVID-19 on survey data quality are then covered.

7. Unfortunately, not all impacts can be assessed ex-post; in fact, many require controlled experiments to truly understand what is behind all the changes. For example, are respondents reacting similarly to the same question if asked through different modes? Was remote training of interviewers as effective as in-person ones? These can be answered only if a controlled experiment is carried out to test the differences. This then points to the response to the second question – how countries can be better prepared for the next crisis. As a key lesson learnt from the pandemic, as highlighted in this Guidance Note, countries should start incorporating bridge studies, experimentation and piloting as a part of their regular survey operations to tease out the real impact of changing data collection mode.

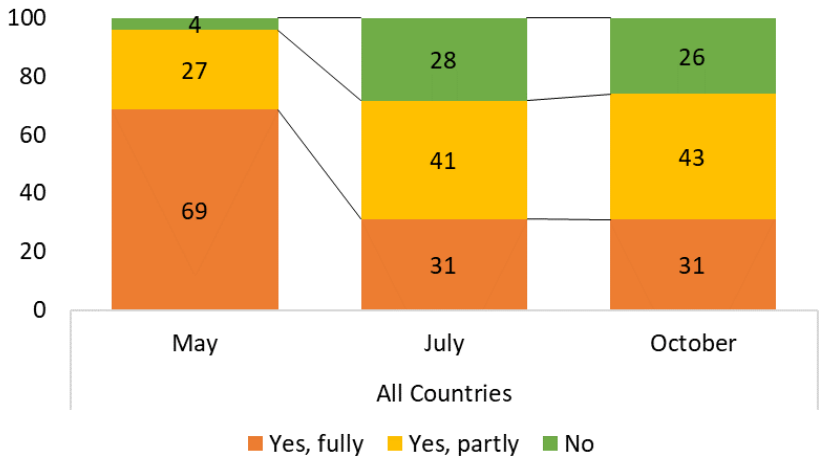
8. Methods covered in this Guidance Note are not new. They have been well-established and tested in many contexts, especially in countries that have started using mixed-mode survey data collection. However, the Guidance Note still provides useful information about national practices during the pandemic and serves as a guide that many countries can relate to.

9. Lastly, the Guidance is very much linked to COVID-19 and rightly so when many countries still struggle to leave the pandemic behind. However, most of the discussions and methodologies covered in the Guidance Note are also relevant in a broader context when survey operations are interrupted by a crisis and need to be adapted. We hope this work helps us in new ways of thinking, planning and using household surveys based on lessons learnt from COVID-19 so we can be more prepared for future crises.

I. Introduction

10. Beginning in March 2020, the coronavirus pandemic urged National Statistical Offices (NSOs) across the world to adjust household data collection activities. According to surveys of NSOs carried out by the UN Statistics Division and the World Bank, 96 per cent of the countries stopped face-to-face survey data collection, either fully or partially, in May 2020 (Figure 1).

Figure 1 NSOs that stopped face-to-face data collection in 2020 as a consequence of the COVID-19 pandemic



Source: United Nations Statistics Division and World Bank. (2020a). Monitoring the State of Statistical Operations under the COVID-19 Pandemic, highlights from the third round of a global COVID-19 survey of National Statistical Offices (NSOs). Available at: https://covid-19-response.unstatshub.org/survey/covid-19-nso-survey-report-3.pdf

11. As a result of the challenges of face-to-face interviewing during the pandemic, many countries adopted innovative approaches to meet the increased data demand. When asked about the changes to be introduced to their planned surveys, more than 50 per cent reported a change in data collection mode or use of alternative data sources; around 40 per cent had added COVID-19 related questions to their surveys and 15 per cent reduced survey length or sample size. (United Nations Statistics Division and World Bank, 2020b).

12. These changes and adaptations introduced to household surveys during the pandemic have the benefit of providing timely data during an emergency situation. However, they could also potentially impact data quality and comparability. Changing the mode of data collection from face-to-face to telephone could introduce biases in the results when the telephone penetration rate is relatively low in the country and those with telephones have different outcomes compared to those without telephones. Incorporating new questions or removing existing questions without thorough testing is likely to affect the original flow of questions and hence the overall survey data quality. Yet changes introduced in data collection mode and questionnaires are an oversimplified overview of what has happened to national household surveys during the pandemic; for instance, changes in training, survey monitoring, and quality control procedures could be introduced. Each of these cascading actions could impact the quality of data provided by NSOs.

13. As pointed out by Discenza and Walsh (2021), reporting on a global ILO survey of impacts of the pandemic on labour force surveys, "it should be noted that some countries did report having concerns about increased bias but were unsure about good methods to assess and correct for it. Thus, there is a need for further support and guidance to these countries, both about bias, but also more generally about the assessment and reporting of data quality to support appropriate interpretation of published results."

14. At the regional level, an example of support is provided by the Statistics Division of ECLAC that promptly acted in 2020 monitoring and assessing the effect of the pandemic on official statistics production. In Latin America, in light of the rapid spread of the disease, governments-imposed curfews, movement restrictions and compulsory quarantines. These mobility restrictions impacted the normal functioning of society, and government institutions were also considerably affected. Because of this situation, the Statistics Division of ECLAC sent out a short questionnaire to the NSO and central banks of Latin America and the Caribbean to learn about the health emergency's effects on the functioning of statistical institutions. Responses from 20 countries indicated that the most affected statistical operations were surveys, administrative records and censuses. Concerning household surveys, most countries claimed that they should postpone information collection because of movement restrictions and the closure of establishments.

15. Also, a series of webinars promoted by ECLAC in conjunction with the International Labour Organization (ILO) and the National Institute of Statistics (INE) of Chile were held to assess the impacts of the COVID-19 pandemic on statistical operations. These webinars were streamed through the Knowledge Transmission Network of the Statistical Conference of the Americas, chaired by Colombia's National Administrative Department of Statistics (DANE).

16. Two years later, after the end of the emergency, the problems of bias, errors, series reconciliation are still present, and include the need to be prepared for other potentially disruptive future events. If we could compare the estimates of surveys carried out during COVID-19 with a pre-COVID survey, the difference would include the impact of all changes introduced to the survey during the pandemic and the real change that occurred during COVID-19 on the variables. Through the Guidance Note, we offer tools to assess the impact of changes introduced to the survey on estimates that could potential enable teasing out the real changes brought by the pandemic.

17. The structure of this Guidance Note is as follows: after a brief introduction, Chapter II summarizes these changes and discusses the challenges and quality issues associated; Chapter III suggests methods for measuring and minimizing their impact; Chapter IV covers a short discussion on the use of

non-probabilistic sampling; and finally, Chapter V makes suggestions on the dissemination of data and metadata for surveys carried out during COVID-19.

18. This Guidance Note is produced under the Inter-Secretariat Working Group on Household Surveys Task Force on COVID-19. Materials were developed based on literature reviews and national practices on survey data quality, measuring and minimizing the impact of changing mode of data collection, changes in questionnaire design and other aspects of survey operations. The Guidance Note also benefited from numerous reports and knowledge-sharing webinars and training sessions at the regional and international level on national practices in household surveys during the COVID-19 pandemic; as well as communications through emails and meetings with a number of NSOs, including Finland, Grenada, Kenya, Niger, Republic of Korea, Singapore, and United States of America about their survey operations during the pandemic and methods used to maintain and improve survey data quality.

19. A comprehensive overview on changes introduced to household surveys during COVID-19 and their potential impact on data quality is presented; however, tools to assess and reduce the impact of COVID-19 on survey quality mainly focus on the impact caused by the changing of data collection mode.

20. While the focus of the Guidance Note is on how COVID-19 impacted on survey operations and quality, the same methodology applies to any other crisis when usual face-to-face interviewing cannot be carried out. We welcome national statistical offices use the guidance to assess their survey quality impacted by COVID-19 as well as in any other similar crisis.

II. Household survey operations pre- and during COVID-19: Potential impact of the pandemic on survey quality

A. Relevant concepts and definitions

21. This subsection covers discussions around key concepts and their definitions that are most relevant to the Guidance Note, including (a) sampling frame; (b) random digit dialing (RDD); (c) error components in household surveys; (d) mode of data collection; and (e) mode effect including the mode-specific selection effect and the mode-specific measurement effect. For each concept, a basic definition is provided, followed by a discussion on the potential impact of COVID-19 on this aspect of household surveys.

22. *Sampling frame.* Any device (lists of areas, addresses, telephones, satellite grids, etc.) that is used to identify and locate units in the population and consequently to select the sample. Ideally the sampling frame should match to the target population. If the sampling frame does not include all of the target population about whom inference will be drawn, then there is coverage error.

23. For a typical face-to-face household survey, the target population is the "non-institutional" population residing in the country, and a frame usually comes from a recent population census or an administrative data source such as a population register. When the data collection mode is postal or telephone, a ready-to-use frame is not usually available for NSOs to select sample directly. In the USA, for example, an address-based sampling frame is built on addresses provided by United States Postal Service. Contact information for telephone or web surveys needs to be acquired from vendors (AAPOR, 2016). During COVID-19, rapid assessment surveys measuring the socioeconomic impact of COVID-19 in countries suffered greatly from the lack of sample frames with telephone contact information (Carletto et al., 2022). Some countries record telephone numbers for households in their census or face-to-face household surveys so they can later use the contact information for telephone surveys. But keeping the frame, especially the contact information, up to date is a serious challenge. In addition, respondents' consent is usually required to use the telephone numbers collected from censuses and face-to-face household surveys for conducting other telephone surveys.

24. *Random digit dialing (RDD) survey.* Telephone numbers in a country are typically structured in a certain way so that a telephone frame can be generated by putting together a list of all possible telephone

numbers that follow the structure. For example, in Ghana, mobile phone numbers are 12 digits, with the first three digits corresponding to the international country calling code for the country (233), followed by two digits corresponding to prefixes for the mobile network operators (MNOs) and the remaining seven numbers are randomly generated (L'Engle et al., 2018). The challenge with the RDD survey is that there are usually a large number of ineligible numbers such as numbers that are invalid, unassigned or belong to businesses rather than households, which usually requires a significant effort for screening. The large proportion of numbers that are of unknown eligibility also requires careful treatment (see

25. Table **3** for the example of an RDD survey in Ghana). RDD started to be used in the United States for official surveys in the 1980s (AAPOR, 2010) but has been losing its popularity due to several reasons, among which high nonresponse rate has been a major element. In low- and lower-middle-income countries, not much was documented on the use of RDD pre-pandemic; hence it is unclear how much experience countries had in using it in surveys carried out during COVID-19. Lastly, an RDD frame usually comes with very little auxiliary information about individual units on the frame, which makes benchmarking and bias correction difficult (see more information in section III.A).

26. *Bias* is a broad concept that refers to the systematic difference between the estimate and the true value if the data collection could be replicated many times on the set of possible samples. Bias can come from different sources. For surveys carried out under COVID-19, the most noticeable sources would be the under-coverage of the sample frame and nonresponse phenomena. Under-coverage and nonresponse alone can cause bias unless proper adjustments are made, especially if respondents differ in the outcome variables from nonrespondents and noncovered population. For example, in Malawi, compared to the pre-COVID-19 face-to-face sample, respondents to the telephone survey were relatively wealthier (Figure 3). The telephone survey is more commonly about the impact of COVID-19 on social and economic outcomes. Thus, if wealthier households were more resilient to the impact of COVID-19 than the less wealthy households, then there will be a bias in the final estimate showing less the impact of COVID-19.

27. *Variance* relates to the randomness on which the statistical inference relies. Under uncontrolled conditions, it can increase and make the estimate unstable, causing larger confidence intervals. Reduced sample size is likely to increase variances. Telephone surveys carried out during the pandemic typically have smaller sample sizes which means variances for survey estimates would be larger compared to large face-to-face surveys pre-pandemic. Also, when weights are used extensively to reduce potential biases, the variance could increase.

28. *Mean square error* of a characteristic x, from a design-based perspective, is a function of bias and variance and can be expressed as $\text{variance}(x) + (\text{bias}(x))^2$.

29. *Mode of data collection.* Survey data can be collected by various modes, including face-to-face, telephone, mail-in (postal), or web. Surveys can use one of the data collection modes or a combination. Surveys that use a combination of survey modes are called "mixed-mode surveys".[1] In these surveys, modes can be applied in different ways. According to Discenza and Walsh (2021), around 50 per cent of about 100 countries had planned to use only the face-to-face mode for their labour force surveys (LFS) in 2020, and 8 per cent of the surveyed countries planned to use remote data collection only. The rest planned to go with a mixture of face-to-face and remote mode. During the pandemic, most countries had conducted at least one remote data collection, mostly through telephone (ISWGHS, 2022a). A web survey is mainly available in high-income countries – one example was the probabilistic web panel survey "Portrait of Canadian Society" carried out by Statistics Canada (Statistics Canada, 2022a). Mixed-mode surveys can take many forms depending on how they are set up in countries. Austria used the face-to-face mode for its first wave of LFS and then a mixture of telephone, web and face-to-face for subsequent waves (Hartleib, Langer, and Moser, 2021).

[1] Mix-mode surveys can be either sequential, or concurrently, more discussions are available in Schouten et al. (2021)

30. For surveys carried out during COVID-19, a large proportion of countries switched the data collection *mode* from face-to-face to telephone. Some countries also used web surveys. Changes in the data collection mode can greatly impact data quality, which is the focus of this Guidance Note.

31. *Mode effect.* The Guidance Note follows the terminology adopted by Schouten et al. (2021) to denote "mode effect" as the compound effect of changing from one mode to another, which measures the effect of changing the mode on the mean square error (variance plus bias). Therefore, the mode effect here covers both the *mode-specific selection effect* and the *mode-specific measurement effect*.

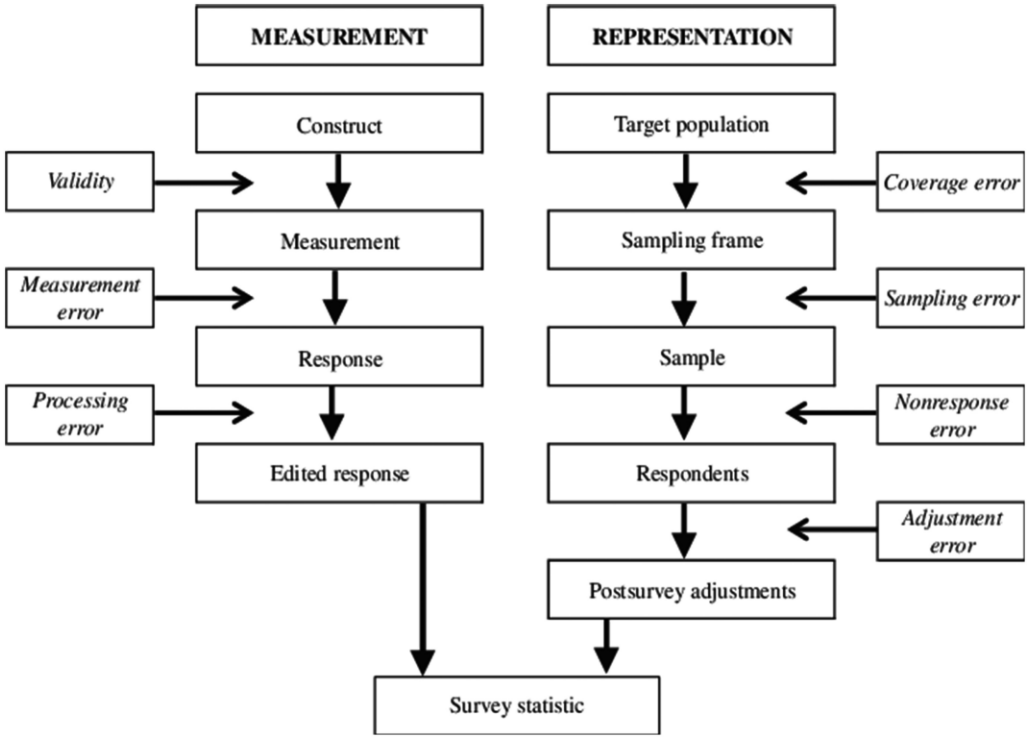
32. *Mode-specific selection effect* refers to the impact of mode changes on representation (Figure 2). For surveys carried out during the COVID-19, this may include a) coverage error such as the incomplete telephone frames compared to the geographic area frame in pre-COVID-19 face-to-face interviews, and b) nonresponse error arising from a higher nonresponse rate for telephone surveys compared to face-to-face survey. When survey respondents are allowed to choose the mode, those choosing one mode may be different in terms of their characteristics from those choosing the other mode. This complicates the ability to assess comparability and contributes to the mode-specific selection effect.

33. *Mode-specific measurement effect* refers to the "pure" mode effect that is commonly used to see how the same respondent responds differently to different modes of data collection. This effect corresponds to the impact of mode changes on the measurement part of the total survey errors (Figure 2).

B. Data quality frameworks

34. Before measuring the impact of COVID-19 on survey quality, it is important to define what survey quality means. The well-known Total Survey Error (TSE) framework (Groves et al., 2004) will be used to illustrate how measures taken during the pandemic could impact the overall survey quality (Figure 2). The TSE approach is very helpful for identifying the many places where data quality can be adversely impacted. If the measurements collected on the questionnaire do not reflect the underlying constructs of interest, the validity of the estimates will be affected. If survey responses do not accurately reflect the measurements that were intended, there will be measurement error. If the sampling frame (list of addresses or telephone numbers) does not perfectly match the target population, about whom inference will be drawn, then there is a coverage error. When collecting data during the pandemic, all seven sources of error identified in the TSE survey lifecycle have the potential to be affected, perhaps substantially.

Figure 2 Survey lifecycle from a quality perspective



Source: Groves et. al. (2004)

35. Additional quality dimensions that are not covered by the TSE framework, including ***timeliness and punctuality, accessibility and clarify, coherence and comparability, and metadata,*** are all important elements for NSOs when changes were introduced to surveys during COVID-19. These are part of the United Nations National Quality Assurance Frameworks (UN-NQAF) for Managing Statistical Outputs (United Nations, 2019), although not a focus of discussions in this Guidance Note, are important considerations for studying the impact of COVID-19 on survey data quality. They will be covered briefly in Session III.C.

C. Impact of changes introduced during COVID-19

36. Before going into details to explain how changes in surveys introduced during COVID-19 relate to the overall quality of household surveys, it is important to note the following three points.[2]

37. First, quality assessment in this Guidance Note is addressed in both absolute and relative terms. In absolute terms an estimate ($\bar{y}$) from surveys carried out during the pandemic is assessed against its true value ($\bar{Y}$), measured through mean square error that covers both the variance and bias. In relative terms the quality of estimates for surveys carried out during COVID-19 will be compared against the "gold standard", which is the pre-COVID-19 survey, for comparability over time and across different population groups. The absolute quality assessment is well covered by the TSE framework (Figure 2), while the relative assessment corresponds to coherence and comparability under the UN-NQAF for Managing Statistical Outputs.

38. Second, the impact of changing mode is meaningful only when referring to specific parameters of interest. Within a single survey, there will be as many "mode effects" as the number of questions. As the

[2] Extracted from Schouten et. al. (2021). Mixed-Mode Official Surveys: Design and Analysis.

Guidance Note includes a significant amount of discussion on the impact of changing mode on data quality, the discussion will be limited to a few parameters for practicality purposes.

39. Third, as covered in the earlier section on basic concepts and definitions, the Guidance Note follows the terminology adopted by Schouten et. al. (2021) to denote "mode effect" as the compound effect of changing from one mode to another. Therefore whenever "mode effect" is mentioned it covers both the "mode-specific measurement effect" and "mode-specific selection effect". Both are defined under section A in this chapter.

40. The pandemic has potentially impacted data quality across a range of components, with particular emphasis on accuracy, timeliness and coherence with previous estimates. The most apparent impacts come from switching data collection away from face-to-face to an alternative mode, whether telephone, internet, or postal. But there are also other factors, such as changes in questionnaire design and lack of proper testing and training that could impact the quality of surveys carried out during the pandemic.

41. The following section discusses in detail, for each type of change introduced in survey design and operation, the potential impact on survey data quality. Such discussion will be linked directly to the data quality framework in section II.B whenever relevant.

i. Changing sampling frame

42. One of the largest potential sources of bias introduced by the pandemic is the loss of complete survey ***frame coverage*** by switching from face-to-face data collection to some other data collection mode. For many countries, household survey samples are usually selected from a frame based on the most recent population and housing census, typically conducted every 5 or 10 years. All housing units have a chance of being included in the sample.[3]

43. While face-to-face data collection theoretically provides every household with a chance of inclusion in a survey, that is not true for other modes of data collection. Depending on the country, a postal address list may be incomplete or out of date, or a large percentage of households may not have access to the telephone or internet. Out of 88 countries and territories with data on mobile phone ownership since 2014, only 48 (55%) reported that 80 per cent or more of their population owning a mobile phone. Only one of the 17 countries in sub-Saharan Africa, with data for the indicator, had mobile phone ownership of 80 percent and above. In 2019, 51 per cent of the world population used the internet and the percentage was 18 per cent in sub-Saharan Africa (United Nations, 2021). Ambel et al. (2021) reported that 99 percent of households in the face-to-face survey have access to a telephone in Nigeria; however, the rate was only between 70 and 80 percent in Ethiopia, Malawi, and Uganda. UNICEF (2021) reported that telephone penetration in Mongolia is 95 percent. Kenya conducted a telephone survey of the socio-economic impact of COVID-19 using telephone numbers from its 2019 census for approximately two-thirds of all the selected households (Ochieng, Z.O., 2021). There is no common standard as to what level of telephone coverage is sufficient for a phone survey – the Gallup World Poll, for example, requires 80% of telephone coverage for telephone surveys (Jong, J, 2016). It is important to note that frame coverage at the national level might be relatively high, but the disparity could be large when zoom into the rural areas or to the most vulnerable population groups. Such disparity within the country needs to be taken into account to avoid excluding certain population groups completely.

44. Reducing frame coverage further is the fact that complete listings of those who do have a telephone, access to the internet, or a postal address are also unlikely to exist and/or be accessible to NSOs. Thus, when face-to-face data collection is replaced, the frame from which to draw the sample is far

[3] Treatment of nomadic populations and others without a usual place of residence is beyond the scope of this document. It is also important to note that most household surveys typically exclude the institutional population which would have an impact on survey estimates.

from complete. The impact on survey quality results from the fact that those who do appear on these frames are not representative of the population as a whole, introducing potential biases. Those with a telephone or access to the internet are likely to be wealthier and better connected to potential services (e.g., health, financial) than those who are missing from the frames (Ambel, et al., 2021); postal address lists are also more likely to be incomplete for small towns, poor and rural areas. In the United States, the National Health Interview Survey switched from face-to-face to telephone data collection and found that respondents skewed towards older and more affluent households (Blumberg, 2021). This may introduce a source of bias to survey estimates not previously of concern.

45. These alternatives are generally less expensive than face-to-face, which allows larger samples to be obtained compared to past data collections. However, even if the sample size increases, the bias will not reduce and the overall mean square error (MSE) is likely to grow. Most importantly, the bias component of MSE is likely hard to measure and thus to communicate to users of the survey data. Data collected from these alternative frames will also be somewhat less comparable to past data than usual measures of change across time. These limitations will need to be communicated as well.

46. Many NSOs prioritized maintaining data collection for panel surveys during COVID-19, while delaying cross-sectional surveys. Typically, the same household is maintained in the panel survey for multiple waves of data collection. While the first wave is generally collected face-to-face, subsequent waves may be conducted by telephone/web/post with the appropriate contact information collected during that first visit. As a result, when the pandemic hit, many NSOs had contact information already available from previous waves, allowing for subsequent data collection via telephone or other modes while face-to-face contacts were not possible. In addition, while a new face-to-face panel could not be introduced during the pandemic, the sample could be supplemented by going back to a completed panel and collecting one more round of data from them. This allowed NSOs to maintain overall numbers of completed cases but did not eliminate loss of frame coverage for new housing units and introduced weaknesses in measuring change over time because these "new" respondents did not remain in sample for the next wave. Another challenge worth noting is that contact information may not be available for all households from the earlier panel, either because some households did not have a telephone, the telephone number was incorrect, or they were not willing to provide the contact information (UNECLAC, 2022). This further aggravated the challenge in frame coverage. Sample attrition is also a common challenge for panel surveys when the households drop out of the survey due to reasons such as respondent fatigue or relocation.

47. The use of RDD was quite popular for countries that did not have access to a reliable telephone frame from other sources. RDD has an advantage over soliciting telephone numbers from an incomplete registration. However, telephone penetration is one thing to consider, as has been discussed extensively above. The other challenge is the lack of geographic information for maintaining the design strata across samples. There is usually no good linkage of mobile phone numbers with the geographic location of the individuals who own the phones. Auxiliary variables are also almost non-existent, which is a challenge for correcting biases after the data are being collected (Chapter III). Another challenge is the high nonresponse rate associated with RDD selection. For example, Tortora (2004) documented a decline in response rate from 1997 to 2003, with only 20.4% of response rate in a large quarterly RDD survey conducted by Gallup. Understanding nonresponse in RDD surveys is more complex than in others because, as previously mentioned, a large number of telephone numbers generated through RDD are ineligible, such as numbers that are invalid, unassigned or belong to businesses rather than households.

48. For countries that used a preexisting list such as population registers, electoral rolls and telephone directories from telecommunication companies for telephone numbers, the coverage of the frame should always be an important element to consider. For example, not everyone is covered by the administrative source and not everyone on the register has an up-to-date telephone number. The high number of relocations or temporary moves occurred during the pandemic intensified the cruciality of this aspect (Frost, 2021).

49. Non-probability sampling has also been used during COVID-19 especially when there is no proper frame to select respondents for remote data collection (ISWGHS, 2022a). This has not been the mainstream in National Statistical Offices given the range of potentially unknown biases associated with the sample. They are exceptions. Statistics Canada, for example, implemented crowdsourcing (non-probabilistic) web surveys to assess the impact of COVID-19. Such data collection is considered essential by Statistics Canada, to obtain citizens' input on their priorities; to collect data in emergency situations such as COVID-19; and to gather information when there is no other source available.[4] It is important to note that not all non-probabilistic surveys are created equal – their quality varies. Designing such survey carefully to ensure inter-operability of the survey with other nationally-representative survey for benchmarking purposes is essential. In the Canada case, a high-quality nationally representative web panel was also deployed. Benchmarking exercise is available comparing estimates from the crowdsourcing web survey and the probabilistic web survey (Beaumont and Rao, 2021)

ii. Changing data collection mode

50. Changing the ***mode of data collection*** from face-to-face to telephone, web, or postal introduces a range of changes in survey quality; in most cases these mode changes made survey quality worse[5], but in a few there were opportunities to improve quality. A few countries (e.g., Korea (Kim, Lim, and Lee, 2021)) only switched part of their data collection, using a mixed-mode approach (Schouten, et al., 2021) combining face-to-face with new modes. Many countries have been using a mixed-mode approach for their Labour Force Surveys (LFS) for a number of years, which was helpful when the pandemic hit (Taskinen, 2021; Tavan, 2021; and Torsteinsen, 2021). In Latvia (Zalkalne, 2021), for 80 percent of LFS respondents in 2017-2019 data was collected using CAPI, with the remainder utilizing CATI or CAWI. In 2020, this flipped to 80 percent CATI, and the remainder CAPI and CAWI. Different parts of the population are more likely to choose specific modes by which to respond. For example, Luxembourg (Schork et al., 2021) found that 20-49 years old were more likely to respond via the web, while those aged 50 and older preferred to answer by telephone.

51. Mode changes can affect quality in different ways. First, as just discussed, the set of respondents can change when the mode is switched; Schouten et al. (2021) refer to this as mode-specific selection effects (versus measurement effects). As seen from the example above in Luxembourg younger individuals respond better to the web mode while older individuals are more likely to respond to telephone. For countries that switched from face-to-face to telephone interviewing, there is generally a lower response rate partly due to unstable telephone connections, or individuals' reluctance to answer to unknown telephone numbers. As recorded by Discenza and Walsh (2021), the cooperation of respondents was a major challenge for countries moving to remote interviewing for the first time.

52. For example, the U.S. National Health Interview Survey (NHIS) saw its response rate drop from around 59 percent before the pandemic to around 42 percent in 2020. Table 1**Error! Reference source not found.** from the NHIS demonstrates the concerns about differential response rates (Blumberg, 2021). First quarter of 2020 response rates were conducted mostly face-to-face and were not strongly affected by the pandemic, but in the second quarter face-to-face interviewing was impossible resulting in very different responses. Respondents were more likely to be 65 or older and less likely to be under the age of 30. They were less likely to be from minority groups or to have not completed secondary school. They were more likely to own a home and have lived there for many years. All of these are highly correlated

[4] More information on the crowdsourcing survey in Statistics Canada on the impact of COVID-19 is available at: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1373185

[5] Some sensitive questions, such as about drug use, are best conducted as self-administered. Respondents may underreport illegal activity to an interviewer. Thus, when an NSO replaces face-to-face with CAWI or postal surveys these questions may achieve higher quality.

with having higher incomes. Weighting was able to eliminate many of these biases, but adults living alone and those in poverty continued to be underrepresented even with weights.

Table 1. Differential response rates from the U.S. National Health Interview Survey (NHIS).

NHIS Sociodemographic Comparisons		
	2020 Q1 (mostly CAPI)	**2020 Q2 (CATI)**
Age group		
18-29 years	**13.6**	**11.1**
65 years and over	**28.8**	**31.5**
Race/ethnicity		
Hispanic	**14.4**	**11.3**
Non-Hispanic black	**11.2**	**9.4**
Non-Hispanic white	**66.7**	**71.9**
Educational attainment		
Less than a high school diploma	**9.5**	**6.7**
BA/BS or higher	**36.5**	**40.5**
Own residence	**65.1**	**73.1**
Years at current residence		
Less than one year	**13.0**	**8.5**
1 to 3 years	**21.6**	**20.0**
4 to 10 years	**24.8**	**24.8**
11 to 20 years	**19.2**	**22.2**
More than 20 years	**21.4**	**24.5**
Total family income		
Less than $35,000	**26.0**	**20.5**
$100,000 or more	**26.6**	**29.2**

Source: Blumberg, 2021

53. The quality concern from lowering response rates is twofold: precision and bias. Precision is reduced if the study ends up having fewer responding households. Lower than anticipated response rates result in fewer completed cases, and therefore less precise estimates. Bias is introduced when respondents are different from nonrespondents. For example, those who are hard to reach may be more likely to be employed away from the home. The biggest concern is that differential response rates by characteristics such as age, sex, or geographical region will introduce biases into many survey estimates. The extent of the bias is likely to vary across outcome indicators, as only some are correlated with the factors that affected response rates.

54. The Korean Household Finances and Living Conditions (HFLC) Survey did not see a noticeable drop in response rates (90.4% in 2019 and 90.1% in 2020) as it switched from primarily face-to-face to mixed modes (face-to-face, self-administered, telephone, and other methods), but the mode of data collection did change substantially. The proportion answering by other than face-to-face mode jumped from 4 percent to 22 percent of respondents (Kim, Lim, and Lee, 2021). Fourteen percent were self-administered, while the other 8 percent responded by telephone and other modes. It was also important to

recognize that these proportions varied across municipalities, with the areas most affected by COVID-19 having practically no in-person responses. Modes also varied by size of household and age of household head. This is an example of how, for items with mode effects, overall response rate can be a misleading indicator of quality. In the HFLC the response rate remained high, but the modes used to collect data changed, and that change was not uniform across the country. It is important for each NSO to not only track overall response rates, but by mode, data collection period and geography to identify potential quality concerns that need to be investigated.

55. The second type of change introduced by changing data collection mode is related to how respondents' answer varies by mode. (Jäckle et al, 2011). This is referred to by Schouten et al. (2021) as mode-specific measurement effects. Even if the same question wording is used, the responses may change from the same respondent. This most frequently happens for sensitive topics such as family planning and drug use. For example, the respondent may not feel comfortable sharing these responses with an interviewer but might do so on a self-administered survey, or the respondent may feel a desire to satisfy what they perceive to be socially acceptable responses, even if they are not truthful. This behavior is likely to be culturally dependent so each country should examine its survey content to identify questions that are potentially subject to mode effects. This can actually work in both directions, some may be comfortable discussing a topic in person, while others may prefer reporting on sensitive topics when there is not a person directly asking the questions.

56. In addition to questions on sensitive topics, questions that allow the respondent to select multiple responses (e.g., "Which of the following…"), might also lead the same individual to different responses. Modes in which the respondent can see all of the choices often obtain a higher average number of selections than when the list of choices is read over the telephone (Jäckle et al, 2011).

57. Changing mode of data collection may also affect other quality-related aspects including: (a) questionnaire design; (b) training of enumerators and supervisors, and instructions for respondents in self-administered surveys; (c) data collection protocols; (d) supervision and quality control of field work; and (e) data entry system. More discussion on changing questionnaire design will be covered in the next subsection, while the other aspects will be covered in subsection iv.

iii. Changing survey questionnaire

58. Under the context of the pandemic, two aspects of changing survey questionnaires are relevant for quality assessment: (a) changing the content of the questionnaire and (b) changing how questions are asked.

59. Respondents typically are more receptive to longer questionnaires in-person than via other methods, thus there is pressure to shorten questionnaires in remote data collection. The necessity to reduce questionnaire length in some countries is also compounded by concerns over telephone connectivity. On the other hand, many countries wanted to add coronavirus-related questions to better assess the impact on households, putting additional pressure on the questionnaire length. Some sections of existing questionnaires were therefore dropped, others had individual questions dropped. Around 50 per cent of the responding countries indicated that they have added questions to their LFS in 2020; and close to 15 per cent added and removed questions (Discenza and Walsh, 2021). Often time, questions or topics removed tend to be those that do not change over a short period of time such as fertility and citizenship participation (UNECLAC, 2022). COVID-19 also seemed to be the right moment for countries to reassess their data collection - one country removed questions for which tabulations were never previously published.[6]

[6] From conversation with a national statistical office.

60. Even questions that remain may need to be restructured, especially for questions that rely on showing physical materials to guide responses (e.g., meal size, show cards) that do not easily transfer to the telephone or other data collection methods (Nicolaas et al., 2011). In addition, questions that are formulated using complicated vocabulary are often difficult to follow during a telephone interview, or difficult for respondents to understand in a self-administered survey. According to the reflection of a CATI interviewer: "*Sometimes questions are so long that when I finish reading them, I don't understand what the beginning is. The respondent – without written text – all the more.*" (Jablonski, 2014)

61. Any changes introduced to survey questionnaires including adding or dropping questions or changing question formulations or wording are likely to impact on survey quality (validity, measurement and processing errors in Figure 2). Keeping questions unchanged while changing the mode of data collection could also be problematic as explained earlier. Thorough testing is required to maintain survey quality. However, it is not clear whether countries carried out such testing during the pandemic, given the tight time frame to produce data and the constraints preventing a full-fledge questionnaire testing which often involves in-person meetings.

62. As an example, new content often undergoes cognitive testing to confirm that respondents understand what the researchers hoped to measure, then new field procedures are part of a pilot test to assure smooth data collection. Historically, cognitive testing occurred in-person (Lessler, Tourangeau, and Salter, 1989); and many countries also carry out in-person pilot tests. The sudden shutdown of face-to-face data collection and the need to collect information on the impact of COVID-19 most likely did not allow opportunity for either of these. Westat (Edwards, 2021) did report using teleconference technology to conduct virtual cognitive testing that was very successful, but this has not been used on a wide scale.

63. Using the TSE framework, these questionnaire-related changes are affecting the left-hand side of Figure 2, including validity, measurement, and processing errors. As pointed out in the UN-NQAF these changes to the questionnaire will also adversely impact the comparability of data collected during COVID-19 versus before and after.

iv. Other changes

64. For those national statistical offices without experience using remote data collection, the pandemic introduced a complex set of challenges related to training, supervision, data collection, processing and quality control.

65. ***Training*** has been greatly affected by the pandemic, although the impacts on quality of changing training practices are hard to measure directly. Before the pandemic, training for data collection in many countries has consisted of initial train-the-trainer sessions followed by trainers conducting a series of in-person training sessions for staff involved in data collection operation throughout the country. When the pandemic struck all of these in-person trainings were no longer possible (ISWGHS, 2020). The quality control procedures to assure high quality training were based on in-person training, including a large number of question and answering, small group discussions, and role playing that are observed by the training staff.

66. During the pandemic, a combination of asynchronous (materials sent out in advance for data collectors to review on their own) and synchronous (live) training were typically provided for more efficient delivery of information. Training conducted over the internet and by review of mailed materials has their specific challenges compared to in-person training sessions. This is because steps available during in-person trainings, especially close monitoring of role-plays and additional re-training for those who were struggling became more difficult on-line. Countries also used a mix of training "modes" to ensure training quality – for its LFS, the Philippines conducted remote training at the national level with the regional trainers, a mixture of remote and in-person at the regional level with the provincial trainers and in-person only for trainings with data collectors (Guillen, 2021).

67. There are, however, two circumstances where revised procedures can improve training. First, rather than conducting train-the-trainer sessions, with the multiple trainers then conducting many in-person trainings, the use of Zoom/Teams allowed for one consistent training to be offered to all interviewers. This reduced variation in training and likely in interviewing as well. An important caveat is that even though it is technically possible to conduct very large virtual trainings, to maintain quality there must be enough training staff to monitor the many small-sized break-out groups that are necessary to refine and confirm training skills. Some countries made effective use of internet break-out rooms and apps such as Padlet to improve their training (Shimizu, 2021). In Grenada (Brizan, 2021), those who obviously struggled during training received additional follow-on training before being allowed to start work. Second, in some countries where the household survey is quite similar from year-to-year, there may have been little training of interviewers (especially experienced ones) in past years. With new procedures mandated by the pandemic, there was a need for all interviewers to receive training, improving quality for this step in the process.

68. From face-to-face to telephone interviewing, the interaction of interviewer and respondent changes. During a telephone interview, there is no body language available to assist understanding. A proper training to enumerators on interviewing strategies specific to telephone interviewing is also extremely important, including how to make proper introduction and build and maintain rapport. Hiring of enumerators should also be reconsidered following culture and social norms on type of enumerators preferred by telephone respondents (Jong, 2016). Such exercise should be incorporated into the broader process of obtaining cooperation from respondents, or as a part of a respondent-centred survey design that aims to promote cooperation from the respondents and improves overall survey data quality.

69. ***Supervision*** of data collection had to be redesigned during the pandemic. The Philippines (Guillen, 2021) continued their LFS with face-to-face interviews while adding alternative methods as well. For the first time, they began weekly monitoring of data collection. Much of traditional supervision around the world historically took place in the field. While telephone systems exist that allow supervisors to monitor telephone data collection and data entry (in both centralized and home-based CATI), countries that did not already have such systems may not have had time and/or resources to introduce such controls (Discenza and Walsh, 2021).

70. Procedures for ***assuring quality*** with remote (e.g., telephone) data collection are different from face-to-face interviewing. For face-to-face surveys, many NSOs use GIS systems built into tablets to confirm that data collection took place in the neighborhood where the respondent lives. This quality control activity is no longer relevant for telephone interviewing. Re-interviews conducted by Kenya (Ochieng, 2021) were able to identify a few cases in Nairobi where data had not actually been collected. If multiple modes were offered, systems needed to be put into place to monitor progress across the modes and track data collection progress. Data entry and processing can also be different in-person versus remote, particularly if paper-and-pencil was used, rather than computer-assisted technology.

71. ***Data collection protocols*** might need to be re-designed, depending on the level of sophistication in the existing CATI systems:

- Integrated systems where the scheduling of interview times and dialing of sampled households are all incorporated in the software, so a CATI interviewer from their home can conduct an interview that mimics the previous computer-assisted in-person interviewing (CAPI).
- Integrated systems like above, but that are dependent on the telephone interviewing being conducted in a centralized telephone center. The interview itself mimics the CAPI experience; however, when interviewers are not allowed to go to the telephone center due to COVID-19, many quality control mechanisms will not be available to monitor home-based interviewing.
- Interviewers conduct the CATI from their home, reading the questions and entering data into the same computer tablet that they would have used for CAPI. Data entry and processing mimic

CAPI, but managing the flow of telephone calling, rescheduling interviews, and other quality control steps will be different for staff who were not previously trained as telephone interviewers.

- Conducting telephone interviews over the telephone and entering the responses on a paper questionnaire. Quality control procedures built into the face-to-face process are missing, with new procedures needing to be developed.

72. Many surveys collect a combination of information at ***household-level and individual-level***. If the goal of data collection is to gather information at the individual level without the use of proxy information (one respondent reporting for all of the household members of question) then telephone surveys will be more problematic. In some households a telephone is associated with an individual, while in others it is shared by multiple members. Regardless, a telephone respondent maybe somewhere away from others in the household when they are interviewed, making it impossible to share the telephone with others. When switching from face-to-face to telephone it is important to consider if one respondent can provide all of the necessary information as well as the quality impacts of a possible higher reliance on proxy response, or higher partial response within households (United Nations Statistics Division, 2020).

73. Some household surveys make use of a particular ***time reference*** such as LFS that refer to a specific one-week time period to obtain employment status. During the first round or two of data collection after switching modes many countries were delayed in their field period, requiring a longer recall period to the reference week. While this might not affect the ability to recall if one worked that week, the ability to remember details such as the number of hours worked may be affected. Alternatively, some countries may have avoided this recall problem by using a floating reference period ("the previous week") which is not consistent across all interviews. In such situations, with a rapidly changing environment such as the local onset of the pandemic, the responses may be a mixture of before or after important events, which will confound interpretation of estimates.

74. In some countries ***data entry*** was not affected by the change in data collection mode. This is because the same tablets can be used for face-to-face data collection or when talking to the respondent over the telephone. In these situations, most quality control checks can still be applied as in the past, for example range checks, skip patterns, and tracking the time from beginning to end of data collection. When data entry procedures are not adaptable to new modes of data collection, for example, switching between paper and pencil surveys and CAPI, entire new software will need to be written to have consistent range and edit checks, along with other forms of quality control.

75. It is important to note that difference observed in estimates during 2020 and 2021 from earlier period is not necessarily a reflection of poor quality or incomparability. For many items collected on household surveys, the true values are likely to have changed in 2020 and 2021. For example, we know that during the pandemic, in many countries thousands (or millions) of workers returned from their place of work to the community where they have family (Biswas, 2021). As a result, household composition during the pandemic is different than before and after. There may be fewer single-person households and more large household groupings. In addition, there were fewer jobs available during this time. The differences shown in data over time might reflect both issues of quality and the dynamic environment during the pandemic. Disentangling the changing true conditions from the data collection changes will be difficult to measure and communicate.

76. In the United States the Bureau of Labor Statistics (BLS) produces monthly estimates of unemployment based on both household and business surveys. During the pandemic there was a large spike in the estimated unemployment rate. While this change might reflect somewhat the ground truth, it appeared that they were biased downward (too low) because so many businesses were short on staff and were not able to respond to the survey in a timely manner. Those most affected by this delay in reporting were those with larger numbers laid off. Careful communication on the data quality to users was important. More information about data dissemination is available in Chapter IV.

77. Training, supervision, data collection and data entry all can impact the quality of survey data. Depending upon how the new procedures compare with the previous ones, mean square errors can either increase or decrease. Although given how quickly the shift had to be made away from face-to-face interviewing, it is more likely that MSEs increased during the pandemic. These changes are sure to reduce comparability of data collected during COVID-19, against pre-pandemic estimates. If an NSO has control values for key totals (either from an administrative source or from a recent census), they can be used in weighting to minimize the bias; but such adjustments are only likely to reduce, not eliminate adverse effects. More information about the potential adjustment is available in Chapter III.

D. Summary

78. As noted in the introduction, most NSOs have switched from face-to-face data collection before 2020, to an alternative method in 2020 and 2021, and some are now going back to face-to-face data collection. The changes and adaptations included those in sampling frames, data collection mode, survey questionnaire and other changes such as training, supervision, data collection protocols, data entry and quality assurance. With all these changes, there is likely to be more error in estimates from 2020 and 2021 than in other years. These errors are a mixture of additional variance and potential biases. Guidance on how to potentially assess and correct the errors is covered in Chapter III, while reflecting on how to communicate these changes to users will be discussed in Chapter V.

III. Assessing quality and methods to reduce and correct errors

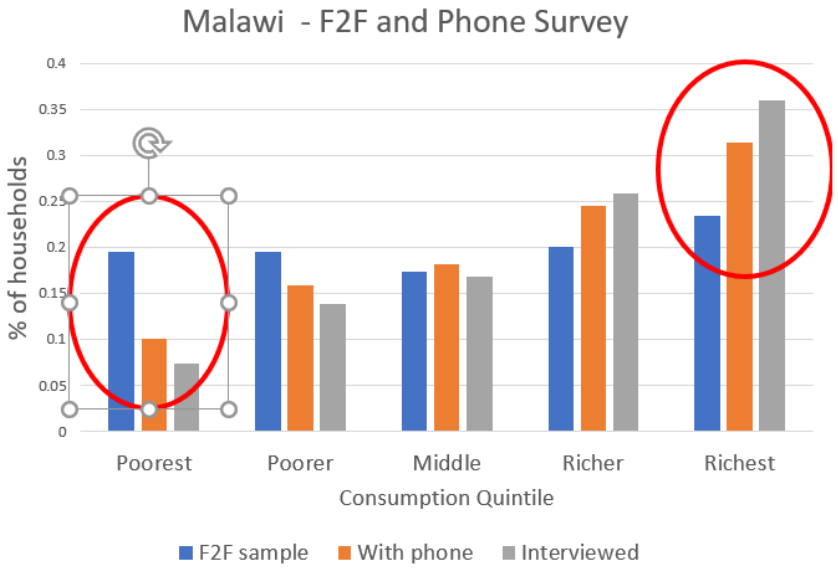
79. As described in the previous section, the coronavirus pandemic has directly impacted the quality of household surveys conducted by NSOs through reduced ***Accuracy*** and ***Reliability***, ***Timeliness*** and ***Punctuality***, and ***Comparability***. We summarize experiences learned from NSOs on each of these components, with a focus on recommendations for measuring and reducing the quality impac In this chapter, guidance is provided on assessing various ways the pandemic could potentially impact household survey data quality; and on methods to reduce the biases or improve survey data quality. It is important to note that not all impacts can be assessed and minimized after data are being collected. Often controlled experiments are needed to tease out the impact. This will be noted whenever relevant.

A. Mode-specific selection effect

80. Mode-specific selection effect is associated with representation errors as shown in the TSE, (Figure 2). As discussed in chapter II, the impact of changing mode on survey representation is linked to sample frame coverage, sample frame list completeness and response rate Estimates are at risk of bias, when households and individuals not covered by the sample frame differ from those covered, in terms of the variable of interest.

81. The following example shows that respondents make-up changed in Malawi after switching from face-to-face to telephone surveys (Carletto, 2020). In the face-to-face sample (blue bars in Figure 3), around 20 percent of the households are in the poorest quintile and less than 25 per cent of the sample are in the richest quintile. The composition of households interviewed in the telephone survey showed that only around 7 per cent were in the poorest quintile and more than 35 per cent in the richest quintile (grey bars). Assuming the face-to-face sample is generally representative, this difference demonstrates that telephone surveys underrepresented the poor, compared to face-to-face surveys (Figure 3). If the outcome indicator of the telephone survey is associated with the level of poverty, then the estimates will be biased. Note that two factors contributed to the bias in the telephone survey: telephone coverage (orange bars) and response rate (grey bar). The blue bars show that the face-to-face sample was spread relatively evenly across the consumption quintiles while the sample was skewed towards richer households in the telephone survey sample.

Figure 3 Change in the make-up of respondents after switching from face-to-face to telephone surveys



Source: Carletto, 2020

i. Calculating response rate

82. Documenting response rate is an important quality assessment step in household surveys. The American Association for Public Opinion Research (AAPOR, 2016) provides standards for the computation of response rates that are used by many countries around the world. Guidance is provided for surveys carried out under different modes of data collection, including mixed-mode ones. If there is a change of mode of data collection for surveys undertaken during COVID-19, difference in response rate calculation needs to be taken into consideration.

83. Table **2** and Table 3 below shows the disposition of a face-to-face Demographic and Health Survey in Ghana (Ghana Statistical Service et al., 2014) and an RDD survey carried out in Ghana (L'Engle et al., 2018), respectively.

84. The largest difference between a face-to-face survey and an RDD survey is in the size of sampling units with unknown eligibility. In the RDD survey, around 28,000 units of unknown eligibility were distributed using the estimated *e*, which is the proportion of cases that are eligible (categories 1 and 2) among all cases with known eligibility (eligible and non-eligible, categories 1, 2 and 4).[7] This proportion will then be used to estimate the number of eligible cases among unknown eligible cases (category 3).

85. From the calculation one can see that the response rate, using different ways of calculation, is between 20-30 per cent for the RDD survey, which is significantly lower than the 95% response rate at household level (and for women) in the face-to-face 2014 Demographic and Health Survey in Ghana (Ghana Statistical Service et al., 2015).

[7] For guidance about how to compute other estimates of *e*, see AAPOR's 2009 Eligibility Estimates. Smith et al. (2009)

Table 2. Sample disposition and nonresponse, Demographic and Health Survey 2014, Ghana

Interview (Category 1)		**Households**
	Complete (C)	11,830
Eligible, non-interview (Category 2)		
	Postponed (P)	0
	Refusal and breakoff (R)	51
	Dwelling not found (DNF)	13
	Household present but no competent respondent at home (HP)	115
Unknown eligibility, non-interview (Category 3)		
Not eligible (Category 4)		
	Dwelling vacant/address not a dwelling (DV)	334
	Dwelling destroyed (DD)	13
	Household absent (HA)	436
	Other (O)	38
Total sample used		12,831
	Response rate at household level (C/C+HP+P+R+DNF)	98.50%

Source: Ghana Statistical Service et al., 2015

Table 3. Call outcomes and response rates, Ghana RDD survey 2017

		Size and calculation
Interview (Category 1)		
	Complete (1.0/1.1)	9,469
	Partial (1.2)	3,547
Eligible, non-interview (Category 2)		
	Refusal and breakoff (2.1)	2,987
Unknown eligibility, non-interview (Category 3)		
	No screener completed (3.21)	1,196
	Unknown if person is a HH resident/ mail returned undelivered (3.3)	27,626
Not eligible (Category 4)		
	Non-working number (4.31)**	918,277
	No eligible respondent (4.7)	2,024
	Other (4.9)**	111,402
Total sample used		1,076,528
	I=Complete Interviews (1.1)	9,469
	P=Partial Interviews (1.2)	3,547
	R=Refusal and break off (2.1)	2,987
	NC=Non Contact (2.2)	0
	O=Other (2.0, 2.3)	0
	Calculating *e**:	0.880
	UH=Unknown Household (3.1)	0
	UO=Unknown other (3.2-3.9)	28,822
Response rate		
	Response Rate 1: I/((I+P)+(R+NC+O)+(UH+UO))	21.1%
	Response Rate 2: (I+P)/((I+P)+(R+NC+O)+(UH+UO))	29.0%
	Response Rate 3: I/((I+P)+(R+NC+O)+e(UH+UO))	22.8%
	Response Rate 4: (I+P)/((I+P)+(R+NC+O)+e(UH+UO))	31.3%

**e* is the estimated proportion of cases of unknown eligibility that are eligible. Enter a different value or accept the estimate in this line as a default. This estimate is based on the proportion of eligible units among all units in the sample for which a definitive determination of status was obtained (a conservative estimate). For guidance about how to compute other estimates of e, see AAPOR's 2009 Eligibility Estimates.

** Eligibility of these two categories is difficult to determine, hence is not included when calculating *e*

Source: L'Engle et al., 2018. Calculation can be replicated using the AAPOR Response Rate calculator 4.1, available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/Response-Rate-Calculator-4-1-Clean.xlsx

ii. Assessing the nature of missingness

86. As seen from the section above, the response rate was significantly lower for the RDD survey compared to the 2014 face-to-face DHS in Ghana. However, higher nonresponse rate does not necessarily mean biases. Whether there is bias in the estimates depends on the nature of the missingness (nonresponse). There are three types of missingness: Missing Completely at Random (MCAR); Missing at Random (MAR); and Not Missing at Random (NMAR).

87. Now borrowing the terminology used in Bethlehem et al. (2011) and the illustration of various types of missingness (Figure 4), let R denote response behavior, which can be seen as a dummy variable, i.e., respond or not to respond; X is a set of auxiliary variables that are available (observed) for all

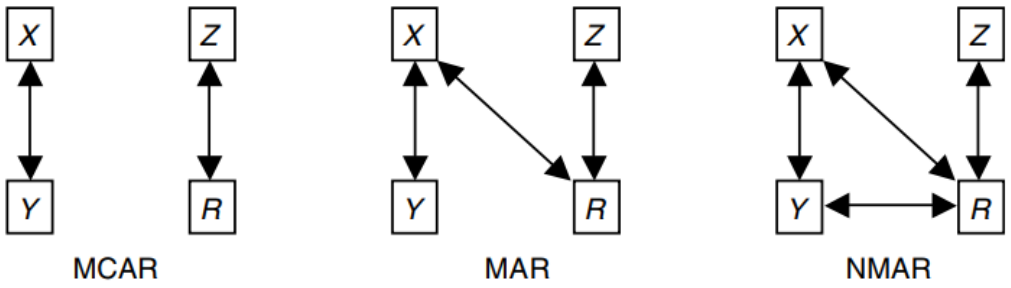
elements within the sample and the population (e.g., age, sex or other demographic characteristics); Y is an outcome variable that the survey is interested in measuring, and finally Z is a set of variables that are partially or completely linked to nonresponse but are unrelated to X and Y. For example, in a traditional LFS, Y can denote the occupational status, X can denote the age, and Z can denote whether the dwelling was occupied in the moment of the collection period.

88. Under MCAR, the outcome variable Y is strongly associated with the auxiliary variable X (employment is strongly related with age) but whether an individual respond or not (R) is unrelated to the auxiliary variable (X) (the response is not related with age). It can be assumed that nonresponse is caused by an independent variable Z other than X (nonresponse is given by the unoccupied dwellings, not by the age of the respondent). In this case nonresponse will not lead to severe bias in the estimates for Y.

89. Under MAR, nonresponse (R) is partially caused by X and partly by an independent set of auxiliary variables Z (for example, nonrespondent rates are higher among those older and lower among younger people). Y and X are strongly related (employment is strongly related with age), and Y is associated with R but indirectly (employment is indirectly associated with age). In this case, the estimates of Y are biased. The bias, however, can be corrected by weighting using the observed value of X (proper adjustments can be made for each group of age).

90. If nonresponse is Not Missing at Random (NMAR), nonresponse (R) is not only related to the set of observed auxiliary variables (X), an independent set of variables Z other than X, but also directly related to the outcome Y (nonresponse rates are higher among those older people that are unemployed). In this case, correction technique with X (weighting) will not help remove the bias in the estimate of Y (any adjustment on age groups will not diminish the bias because older people are currently unemployed and at the same time they are not answering to the survey).

Figure 4. Missing data mechanism illustration



Source: Bethlehem et. al., 2011, Chapter 2, Figure 2.2. Handbook of Nonresponse in Household Surveys

91. Note that even though the language used above refers to survey response and nonresponse, a similar argument also applies to sample frame coverage and noncoverage. If we link back to Chapter II on the impact of changes introduced during COVID-19, particularly the impact of changing sampling frame and data collection mode, the selection bias is a potential issue for both. Under the sampling frame coverage issue, instead of being defined as response/nonresponse, R would be defined as indicating whether a unit (household or individual) is being covered and not covered by the frame.

92. What kind of missingness do changes of surveys during COVID-19 introduce? With the example in Figure 3, one can see that both the sample frame coverage and response behavior (R) have been affected by switching from face-to-face to telephone interviewing in Malawi. If we treat poverty quantile as an auxiliary variable (X), it is associated with R. In the example, households in lower wealth quantiles were less likely to be covered by the telephone survey. If households in lower wealth quantiles behave differently, in terms of outcome indicators (e.g., food security, health status, etc.), from those in the higher wealth quantiles, then there is a bias in the estimates for the outcome indicators. Most likely the

missingness would qualify as Missing at Random (MAR) and the auxiliary variables X can be used to reduce the biases.

iii. Detecting a bias

93. As explained in the previous subsection, biases can arise for certain types of missingness (MAR, NMAR) and can be corrected for MAR. Biases can be detected through testing of independence of the response (and/or coverage) or comparing the difference between respondents and nonrespondents, which is illustrated with a country example (more information about the survey data is in Box 1). Different approaches are illustrated in this section, including (a) bivariate testing of independence; (b) bivariate graphs; and (c) multivariate analysis.

Box 1. Malawi High Frequency Phone Survey: background

High Frequency Phone Survey (HFPS) is being carried out in Malawi as part of the World Bank initiative in supporting countries to measure the socioeconomic impact of COVID-19. The sample frame of HFPS in Malawi was drawn from the 2019-2020 Integrated Household Panel Survey (IHPS). Within IHPS, telephone numbers were collected for up to four members of each household, allowing for the implementation of HFPS.

Bivariate test of independence

94. Bivariate tests of independence can be carried out between the auxiliary variables and the coverage (or response) indicator. As shown in

95. Table **4** below, a two-way frequency table is produced consists of two variables: cellphone ownership at the household level (R) and literacy status of head of household (X). In this particular survey (Box 1), a question was asked whether each household member (10 years and older) has a working cellphone. Households with at least one cellphone could potentially be considered as in the frame for the High Frequency Phone Survey that measures the impact of COVID-19. Whether a household has a mobile phone is considered as a "response/nonresponse" indicator (R) while the literacy status of the household head is the auxiliary variable (X) that is available in both IHPS and HFPS.

96. Pearson's Chi-squared of goodness-of-fit between the two variables of interest (mobile phone ownership and literacy) can be performed to determine if a significant correlation exists. The null hypothesis claims that the two variables are independent, while the alternate at the two variables are not independent. From this data, the test statistic is 332, and the p-value is very close to zero, indicating a strong evidence to reject the hypothesis of independency.

97. Cramér's V test is used to test the strength of association between two categorical variables. In our example, the value of the Cramér's V is 0.329, which indicates moderately strong association between household's ownership of cellphone and literacy status of the head.[8] The association is considered moderately strong. Households with illiterate heads are less likely to have a mobile phone and therefore less likely to be covered by the High Frequency Phone Survey. If these households provide different answers to survey questions in HFPS, there will be bias caused by sample frame coverage.

[8] For k=2 and degree of freedom = 1, [0.1,0.3) indicates small association, [0.3, 0.5) medium association and [0.5,1] large association. More information is available at Manigiafico, S. S. (2016). The degree of freedom is defined to be the minimum between the number of row minus one, and the number of columns minus 1.

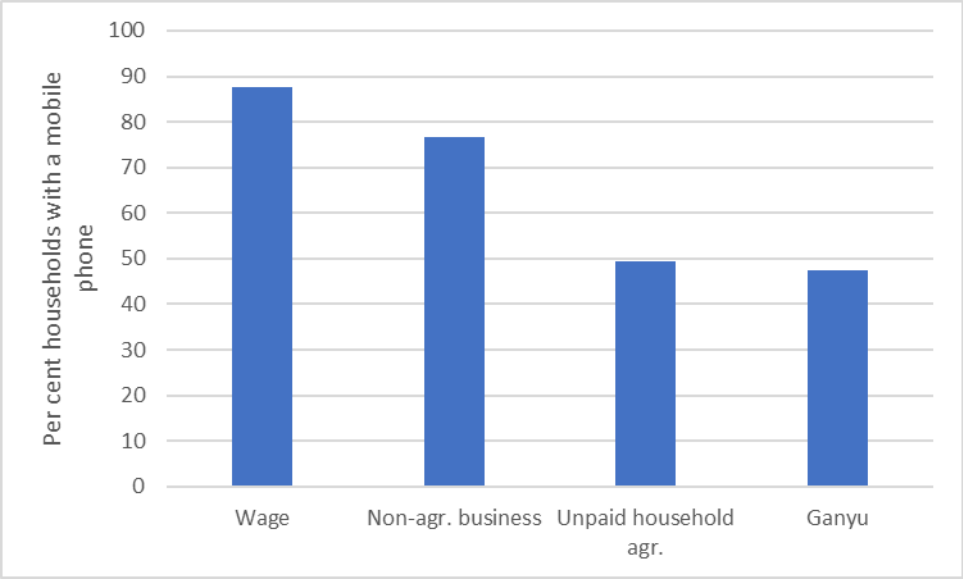
Table 4. Two-way frequency table of ownership of telephone in the household and literacy status of head of household, Integrated Household Panel Survey, 2019-2020, Malawi

		Literacy of household head		
		Literate (%)	Illiterate (%)	Total
Ownership of mobile phone	Yes	1624 (52.8%)	146 (4.7%)	1770
	No	854 (27.8%)	453 (14.7%)	1307
	Total	2478	599	3077

Bivariate graphs – test of independence

98. Visualization tool is also powerful in understanding the relationship of auxiliary variables and the sample frame coverage (or the response likelihood). Figure 5 below illustrates how the mobile phone frame coverage varies by the employment[9] of the household head. This figure clearly shows that household with the head having a wage employment is more likely to have a mobile phone (85%), hence be included in the sample frame for the High Frequency Phone Survey. If households whose head have a wage employment are impacted differently compared to those with heads working as unpaid household worker in agriculture, then there is a bias in measuring the impact.

Figure 5. Proportion of household owing a mobile phone, by employment of household head, Integrated Household Panel Survey, 2019-2020, Malawi.



Note: For illustration purpose, the category "unpaid apprenticeship" is excluded from the figure as it only has 4 cases out of 3,077.

Multivariate analysis: test of independence

99. Bivariate analysis suggested earlier does not take into consideration relationship among different variables. Multivariate analysis model could be constructed based on the bivariate tests done in earlier steps. Stepwise approach can be used to select the model. Using the sample example data, a logistic

[9] The types of employment considered are: 1) employment on wage; 2) working in non-agricultural business; 3) unpaid work in household agriculture; 4) Ganyu, i.e. informal for short-term rural labour; 5) unpaid apprenticeship. Note for illustration purposes, the category "unpaid apprenticeship" is excluded from the analysis as it has 4 cases out of 3077.

regression was fit using two independent variables: literacy and employment status of household head (Table 5**Error! Reference source not found.**). A logistic regression model can be fitted as follows:

$$logit\,(\rho) = ln\,(\frac{\rho}{1-\rho}) = \beta_0 + \beta_1 I_{illiterate} + \beta_2 I_{non-agr\ business} + \beta_3 I_{unpaid\ household\ agr} + \beta_4 I_{Ganyu}$$

where β_i is the coefficient and I_i is the value of the independent variables (literacy status and employment status of household head in this example), and ρ represents the probability of response. In this particular example, ownership of mobile phones is used as a proxy for whether responding or not.

100. As shown in the fitted model (Table 5), both variables are strongly associated with household ownership of mobile phones. The coefficient β_1 for the literacy variable is -1.4954, which means that compared to households with heads who are literate, the estimated odds ratio of owning a mobile phone for households with an illiterate head is $e^{-1.495} = 0.224$. Similar conclusion can be made for employment status of household head. Hence there is a potential bias associated with the telephone sampling frame when using IHPS as a frame for the High Frequency Survey in Malawi because less literate and non-wage employed household heads are less likely to respond and be included in the IHPS.

Table 5. Logistic regression model for household ownership of mobile phone, Integrated Household Panel Survey, 2019-2020, Malawi.

Variable	Category	Coefficient		Odds ratio
Intercept	Overall	1.8712	*	
Literacy status of head of household *Reference: literate*	Illiterate	-1.4954	*	0.22
Employment status of household head *Reference: wage employment*	Non-agriculture business	-0.6236	*	0.54
	Unpaid household agriculture	-1.7658	*	0.17
	Ganyu	-1.8898	*	0.15

Note: * p-value <0.001

Comparing sample and population

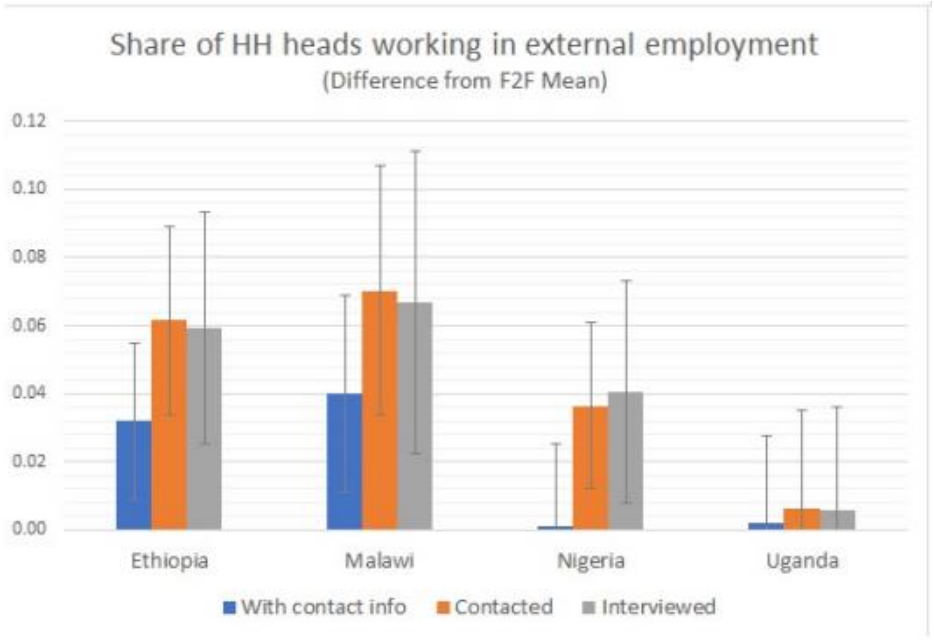
101. Another way to assess bias in sample coverage is to compare the distribution of sample units with the population distribution for a set of auxiliary variables. As shown in

102.

103.

104. **Figure 6** below, the proportion of households with heads working in external employment (having a wage employment) is compared against the benchmarking survey (LSMS-ISA) and standard errors are shown in the same chart. Taking Ethiopia as an example, compared to the LSMS-ISA (face-to-face) estimate, households with contact information has 3 percentage higher in the share of household heads working in external employment. Among households that were successfully contacted for the high frequency survey, the share with the head in external employment is 6 percentage higher than the original LSMS-ISA sample. There was not much difference in the average percentage of head working in external employment when the interviewed households and those contacted successfully were compared.

Figure 6. Share of household heads working in external employment, compared to face-to-face surveys (LSMS-ISA)



Source: Ambel et al. (2021)

105. A multivariate analysis can also be done to compare the sample and the population, for a number of auxiliary variables. As shown in Table 6 household head is more likely to respond, increasing the probability by 39.7%. Individuals who own a mobile phone are more likely to be in the HFPS sample than those who do not own, with a 15% increase in probability. (Brubaker et al., 2021)

Table 6. Marginal effects from logit regressions on being a HFPS respondent in round 1, Malawi

Variables	**Marginal effect coefficient**
Household size	-0.013(.002)***
Head	0.397 (0.026)***
Spouse of head	0.140 (0.033)***
Child of head	-0.006 (0.026)
Male	-0.040 (0.015)***
Ages 25-49	0.040 (0.016)**
Ages 50+	0.038 (0.019)**
Married	-0.021 (0.019)
Consumption quintile 5	-0.048 (0.021)**
Individual owns a mobile phone	0.154 (0.012)***
…	…

Note: ***/**/* denote statistical significance at 1/5/10 per cent level, respectively. The table is shortened from the original version and is included only for illustration purposes.
Source: (Brubaker et al., 2021)

Nonresponse and representativity

106. Survey response rate is a poor indicator of the effect that nonresponse can have when analyzing survey data, especially if we intend to identify biases. A low response rate is undesirable and has a direct effect on the precision of survey estimates (the lower the effective sample size, the larger the variance of the estimator); however, it does not necessarily imply bias. If no relation exists between the response mechanism and the variables of interest in the survey (MAR or MCAR mechanisms in Figure 4), the classical weighting approach will yield unbiased estimates.

107. Representativity indicators (R-indicators) have been defined to measure to what extent the nonresponse in a survey can affect the quality of the estimators and increase the risk of possible biases (Bethlehem, et. al., 2011). For Bethlehem, et. al. (2009) the concept of representativity indicates the absence (or presence) of selective forces on the response mechanism (three scenarios in Figure 4). From a probabilistic perspective, there will not be selective forces on the response if the probability of response (propensity score) across the population units are all the same; that is:

$$\rho_i = Pr(r_i = 1 | s_i = 1) = \bar{\rho} \qquad for\ i = 1, 2, \ldots, N.$$

108. In the above expression, ρ_i is the propensity score of unit *i*; r_i is a dummy variable that indicates whether unit *i* is a respondent, and s_i is a dummy variable that indicates whether unit *i* was included in the sample. If this equation holds, then the response mechanism is called strongly representative.

109. According to Bethlehem, et. al. (2009), the response to a survey is called weakly representative with respect to the sample for auxiliary variable X if the average response probability is the same in each subgroup; this is the case of the MAR in Figure 4. This way, weak representativity means that, given X, it is not possible to distinguish respondents from non-respondents. In this case, biases of estimators will be small.

110. Assuming that individual response probabilities are known for the population, the R-indicator is defined as follows:

$$R(\rho) = 1 - 2\,S(\rho)$$

where $S(\rho)$ is the standard deviation of the response probabilities. Notice that $S(\rho) = 0$ if the response mechanism is strongly representative; also, the larger the variation on the response probabilities, the lower the representativity induced by the response mechanism. By the dichotomous nature of the r_i variables, the maximum value of $S(\rho)$ is 0.5. This way, the R-indicator ranges in the interval *[0, 1]*. A value of $R(\rho) = 1$ implies strong representativity, while a value of $R(\rho)$ close to zero implies that the response composition deviates from the effective sample composition, indicating a high risk of selection bias.

111. As response probabilities are not known, we have to estimate them by means of a working model such the one defined in paragraph 85, using covariates that were observed for both respondents and nonrespondents. Provided that w_i is the sampling weight of unit *i* in the sample, the R-indicator is estimated as follows:

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{\hat{N} - 1}\sum_{i=1}^{n} w_i \left(\rho_i - \hat{\bar{\rho}}\right)^2}$$

where ρ_i represents the response probability for unit *i* and $\hat{\bar{\rho}}$ is the weighted average of responses probabilities in the sample:

$$\hat{\bar{\rho}} = \frac{1}{\hat{N}} \sum_{i=1}^{n} w_i \rho_i$$

112. From the model fitted in the above paragraph, we have that the unweighted response rate was 0.57, while the weighted response rate was 0.53; also, the unweighted response propensity was 0.57, while the weighted counterpart was $\hat{\bar{\rho}} = 0.55$. This way, the R-indicator was estimated as $\hat{R}(\rho) = 0.56$, which is lower than the ideal value of 1.00. Then, this response structure is not completely representative, and more work should be done for avoiding the possible bias that lack of representativity can generate in the analysis of this survey data.

113. Under the working model, the nonresponse mechanism can be considered as MAR and is governed by covariates *literacy* and *employment status of household head*; that means weak representativity. For instance, Table 7 shows that when grouping by literacy, the R-indicators are higher than the one estimated at the national level, indicating higher homogeneity within literate and illiterate individuals; from Table 8, representativity increases when considering Activity. Finally, when considering both covariates, the standard deviation of the estimated response probabilities is null, and estimated R-indicators take the maximum value in all the subgroups (Table 9). This example clearly states that nonresponse rate should be complemented with R-indicators to have a clearer view of the impact that nonresponse mechanism can have in the analysis of complex survey data.

Table 7. Estimated representativity indicators for Literacy

Literacy	N	Weighted response propensity	Standard deviation of propensities	Estimated R-indicator
Literate	2474	0,85	0,06	0,87
Illiterate	599	0,64	0,16	0,69

Table 8. Estimated representativity indicators for Employment status of household head

Employment status of household head	N	Weighted response propensity	Standard deviation of propensities	Estimated R-indicator
wage employment	626	0,85	0,06	0,87
Non-agriculture	584	0,73	0,11	0,77
Unpaid household agriculture	1340	0,43	0,15	0,7
Ganyu	523	0,41	0,13	0,72

Table 9. Estimated representativity indicators for both Literacy and Employment status of household head

Literacy	Employment status of household head	N	Weighted response propensity	Standard deviation of propensities	Estimated R indicator
Illiterate	wage employment	31	0,59	0	1
Illiterate	Non-agriculture	70	0,44	0	1
Illiterate	Unpaid household agriculture	374	0,2	0	1
Illiterate	Ganyu	124	0,18	0	1
Literate	wage employment	595	0,87	0	1
Literate	Non-agriculture	514	0,78	0	1
Literate	Unpaid household agriculture	966	0,53	0	1

Literate	Ganyu	399	0,5	0	1

Assessing bias at more disaggregated level

114. Examples above have shown various ways of assessing potential coverage or sample selection bias. All examples have so far been focused on national level. Coverage could in fact vary greatly by region, sex, age and other subpopulations – therefore further disaggregated investigation should be done to the extent possible.

Summary

115. The subsection covers different ways, with national examples, that countries can use to assess whether there is a potential bias for surveys carried out during COVID-19 due to (a) sampling frame coverage deficiency as discussed in Section D in Chapter II and (b) lower response rate. Non-coverage and nonresponse errors share many similarities in terms of definitions, measurement, reduction, compensation and reporting (United Nations, 2005). However, it is important to note that these two are also quite different in terms of sources of errors and in some case, solutions. For example, non-coverage error in the context of COVID-19 was mainly caused by incomplete mobile or regular telephone coverage (for telephone surveys) and lack of proper frame. Nonresponse could potentially be caused by noncontact (due to change of telephone numbers), break-off (unstable telephone connections) or reluctance of households to respond to a unknown telephone numbers. Methods to detect the biases are similar as outlined in this subsection but there are differences in the availability of benchmarking data sources. Whenever relevant those differences will be highlighted.

116. In Ambel et al. (2021), the assessment of potential biases in the sample were done in two steps with the first assessing sample representativeness by comparing the population (in the Malawi example the face-to-face IHPS) and those with contact information; and the second step assessing nonresponse bias by comparing the selected sample (with mobile phone) and the respondents.

117. Assessing and correcting survey biases depend on a key element which is the availability of the benchmarking data source; and the auxiliary variables that are available in surveys carried out during COVID-19 and in the benchmarking data source. More discussion will be provided in the next subsection.

iv. Selecting benchmarking sources

118. This subsection starts with a general discussion on data sources and auxiliary variables that can be used to assess and reduce potential biases, followed by specific discussions on different approaches used for selecting samples during the pandemic, such as selecting from a recent panel survey, an administrative source, or RDD.

119. Data sources that are commonly used for benchmarking include (a) population censuses; (b) administrative data sources such as population registers; (c) a recent probabilistic survey that was carried out pre-COVID-19. A number of factors need to be considered when searching for appropriate benchmarking data sources. The first is the quality of the benchmarking data source itself. Population census is designed to cover the entire population but may be outdated or come with coverage and measurement errors. Administrative data sources can have their own quality issues such as coverage and measurement errors as well as inconsistencies in concepts, definitions, classification and so. Household surveys, of course, also have sampling and non-sampling errors. Before using these data sources for benchmarking it is important to understand their quality and how the quality issues will impact on the final estimates.

120. In addition to considerations of types of data sources for benchmarking, the availability of auxiliary variables for calibration is also important. To derive weights to reduce the biases due to under-coverage, relevant auxiliary variables need to be available for both the covered and non-covered

populations. Similarly for correcting the biases due to nonresponse, the auxiliary variables need to be available for both the respondents and nonrespondents.

121. The availability of auxiliary variables varies across data sources: administrative data usually offer limited number of variables; censuses have moderate number of variables and household surveys usually come with a large number of variables that can potentially be used for benchmarking.

122. While in terms of non-coverage, surveys almost never know anything other than the location and general characteristics of the non-covered portion of the population, in non-response they know at least frame information for non-respondents. (United Nations, 2005)

123. Common social and demographic characteristics that are used to correct the noncoverage and/or nonresponse bias could include: sex, age, race/ethnicity, marital status, household size and composition, education, home ownership, income, labour force status and location. However, the prediction power of auxiliary variables to the outcome indicators is another important element in reducing biases due to noncoverage and nonresponses. For example, individuals aged 50 years and over are more likely to be in the Malawi high frequency phone survey (HFPS) than those under the age of 50. (Table 6) But is the respondent 50 years and over in the Malawi HFPS like all the others not in the survey who are 50 years or above? Or does the person differ, for example, in telephone ownership or employment status (those not employed are more likely to take a call)? If this is true, nonresponse and noncoverage adjustment carried out only using the variable age is not going to remove the biases. In a study undertaken to study the nonresponse in online and face-to-face survey, personality of individuals was found to be closely linked to whether they respond to online or face-to-face survey (Valentino et al., 2021). Online respondents are less extraverted than face-to-face respondents and such difference impacts on political preference, which is the outcome indicator for the study. The worry then comes as there is no benchmarking source to provide the auxiliary variable on personality. The UK LFS acknowledged the importance of weighting to account for biases for its online only data collection during the pandemic but has proposed an increase in sample size to ensure more inclusive coverage (Morgan, 2022).

124. Depending on how samples are selected, benchmarking data sources and availability of auxiliary variables vary. A panel survey that selects respondents from a recently completed panel (Table 10) could use the previous round of survey (Column B) as benchmarking data source. A large number of variables are available for elements that are not covered in the frame due to lack of contact information and also for nonrespondents. A further calibration with population census that contains information about the target population (Column A) could be carried out after the initial step.

Table 10. Illustration of a new panel survey carried out during COVID-19, based on a recently completed panel

Target population	**Respondents from the most recent panel**	**Households with contact information**	**Households selected (assuming all with contact information are selected; could also be a random sample)**	**Responding households**
A	**B**	**C**	**D**	**E**
100001	X	X	S	R
100002	Not covered in the most recent panel			
100003	X			
100004	X	X	S	
100005	Not covered in the most recent panel			
100006	X	X	S	R
…				
110000	X	X	S	R
X denotes if the household in the target population responded in the most recent panel and has contact information available;				

S indicates the household was selected from the target population having contact information available; **R** indicated the household responded to the survey

125. For cross-sectional surveys that obtained sample from an administrative list (Table 11), column A represents the target population and column B is the administrative list that reflects a common scenario that has a percentage of underrepresentation. Not all individuals in the target population are covered by the administrative source. Column C represents individuals that are covered by the administrative list and have contact information for the survey. A sample (Column D) is then selected from those with contact information and respondents are in column E.

126. It is difficult to tell how many auxiliary variables are available for benchmarking as administrative data sources differ from each other. A population register typically covers information on geographic location, age, sex and other basic demographic characteristics but voter registration probably does not have as much information. Benchmarking can be carried out either through comparing information on all respondents (Column E) and a data source containing information about the target population (Column A); or in two steps – comparing Columns E and B; and then calibrating with data in Column A.

Table 11. Illustration of a cross-section survey carried out during COVID-19, with its sample obtained from an administrative source

Target population	**Administrative list (voter registration, population register etc)**	**Individuals with contact information**	**Individuals selected**	**Respondents**
A	**B**	**C**	**D**	**E**
100001	X	X	S	R
100002	X	X	S	
100003	Not covered on the list			
100004	X	X		
100005	X			
100006	Not covered on the list			
…				
110000	X	X	S	R
X denotes if the household in the target population responded in the most recent panel and has contact information available; **S** indicates the household was selected from the target population having contact information available; **R** indicated the household responded to the survey				

127. For countries that selected sample through random digit dialing (RDD), only basic calibration can be carried out against a source that has information about the target population (such as a population census, column A) (Table 12). As discussed earlier, limited availability of good auxiliary variables poses challenge in reducing the selection bias. Basic demographic characteristics could be used for calibration but might be "cosmetic" if those strongly associated with the outcome indicators are not available.

Table 12. Illustration of an RDD survey carried out during COVID-19

Target population	**Eligible individuals (also with a mobile phone)**	**Respondents**
A	**B**	**C**
100001	X	R
100002	X	
100003		
100004	X	R
100005		
100006	X	
…		
110000	X	R

v. Correcting the bias with weighting and calibration

128. Mode-specific selection bias correction can be carried out using the traditional weighting and calibration methods that have been adopted often by NSOs. This section describes various methods that have been used to reweight the final results, with the intention to reduce noncoverage and nonresponse biases.

Weighting class adjustment

129. The weighting class adjustment method assumes that units, respondents and nonrespondents, within the same class, defined by certain characteristics, have about the same response probability and about the same outcome values. Under this assumption, nonresponse bias will be eliminated after adjustment. However, the outcome values of nonrespondents are not observed and one set of formed classes tends to impact the outcome values differently. Therefore, weighting classes are usually formed with the intention to contain units with the same response probability.

130. As shown below, if we denote the original unadjusted weight as w_i^0 for the i^{th} unit (including both the respondents and nonrespondents) in subclass *c*, then the adjustment for respondents within class *c* is calculated as the total weight for all units divided by the total weight for all respondents within class *c*.

$$a_c = \frac{\sum_{i\in\{all\ eligible\ units\}} w_i^0}{\sum_{i\epsilon\{eligible\ respondents\}} w_i^0}$$

131. After this adjustment, the weight for respondent *i* in class *c* is calculated as $w_i^0 \times a_c$ and the weight for nonrespondent in class *c* is *0*. This way, the new set of adjusted weights for respondents is given by:

$$w_i^{ac} = w_i^0 \times a_c$$

Notice that the population size will be underestimated if we use the unadjusted set of original weights. That is $N \gg \sum_{i\epsilon\{eligible\ respondents\}} w_i^0$. However, once the weights are adjusted, we can expect that $N \cong \sum_{i\epsilon\{eligible\ respondents\}} w_i^{ac}$.

132.

133.

134.

135. **Table 13** below illustrates calculation of a_c using the same survey data as in Table 5**Error! Reference source not found.**. Households in the 2019-2020 Malawi IHPF are classified into 8 classes[10] defined by two variables: literacy and employment status of household head. Columns 3 and 5 represent the number of households covered by each of the classes, for respondents and nonrespondents, respectively. Column 4 is the sum of household weight for respondents in the respective class and column 6 is the sum of household weight for nonrespondents. Column 7 (a_c) is calculated as (column 4 + column 6)/column 4. As shown in the table, higher proportion of nonrespondents produces a higher adjustment

[10] The category "Unpaid apprentice" is not included in the demonstration because of its small sample size.

factor – a very high adjustment factor is usually not a welcoming sign.[11] For example, in the last row for households with illiterate head who works as Ganyu (short-term rural labour), the adjustment factor is 2.94. This is because out of 124 households in this class, more than 80 percent are nonrespondents.

Table 13. Calculation of nonresponse adjustment factor using weighting class adjustment, Integrated Household Panel Survey, 2019-2020, Malawi [12]

Literacy of household head	Employment status household head	Respondents		Nonrespondents		
		# households	Sum of weights	# households	Sum of weights	
Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*
Literate	Wage worker	522	493615.79	73	64019.71	1.13
Literate	Non-agriculture business	398	377020.176	116	99282.32	1.26
Literate	Unpaid household agriculture	503	629570.333	463	626815.89	2.00
Literate	Ganyu	198	176109.941	201	194356.67	2.10
Illiterate	Wage	12	9871.247	19	32073.86	4.25
Illiterate	Non-agriculture business	32	31590.948	38	57437.76	2.82
Illiterate	Unpaid household agriculture	80	117616.305	294	433297.08	4.68
Illiterate	Ganyu	22	64776.812	102	125419.06	2.94

136. Once calculated, the adjustment factor is multiplied to the household weight for responding households (Table 14).

137.

138. Figure 7 shows the relation between the unadjusted weights and the ones adjusted by class. As expected, under the working model, this adjustment yields to weights that reproduce more accurate

[11] When weights are adjusted to deal with different issues such as non-coverage and non-respondence, these adjustments may result in very large variation in the sampling weights which can significantly increase sampling variances. In such cases, NSOs may impose a trimming strategy for excessively large weights. This may lead to an increase in bias and decrease in the sampling variance. However, the objective of trimming weights is to reduce the MSE.

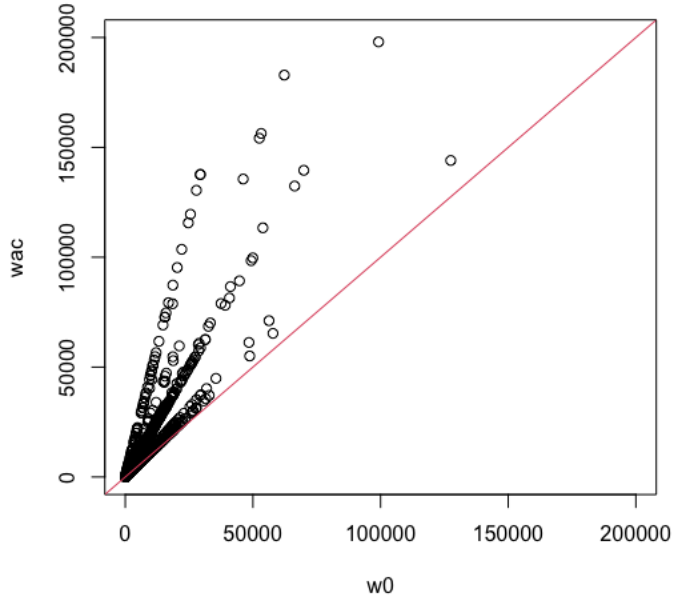
[12] In this example, household ownership of mobile phone is used as an indication for response/nonresponse as data to the actual response behavior is not available.

estimates. Therefore, the underestimation problem is solved, and consequently the new weights are above the 45-degree red line.

Table 14. Adjustment of household weight using nonresponse adjustment factor, Integrated Household Panel Survey, 2019-2020, Malawi

Household ID	Response/nonresponse status	Literacy of household head	Employment status of household head	household weight	Adjustment factor	Weight following nonresponse adjustment
0001-002	Nonrespondent	Illiterate	Unpaid household agriculture	2877.0	-	-
0001-005	Respondent	Literate	Non-agriculture business	2877.0	1.26	3625
0002-001	Respondent	Literate	Unpaid household agriculture	2977.7	2.00	5954
0002-005	Nonrespondent	Literate	Unpaid household agriculture	2977.7	-	-
0002-006	Nonrespondent	Literate	Non-agriculture business	2977.7	-	-
0003-003	Nonrespondent	Literate	Unpaid household agriculture	2977.7	-	-
0004-001	Respondent	Literate	Unpaid household agriculture	3056.7	2.00	6113
0004-005	Respondent	Literate	Unpaid household agriculture	3056.7	2.00	6113
0029-003	Respondent	Literate	Wage worker	3184.9	1.13	3599
…						

Figure 7. Class variable adjusted weights and the original unadjusted sampling weights



Note: The 45-degree red line represents equality of unadjusted sample weights (w_i^0, in x-axis) and weights after the simple class-variable adjustment (w_i^{ac}, in y-axis)

Propensity score adjustment

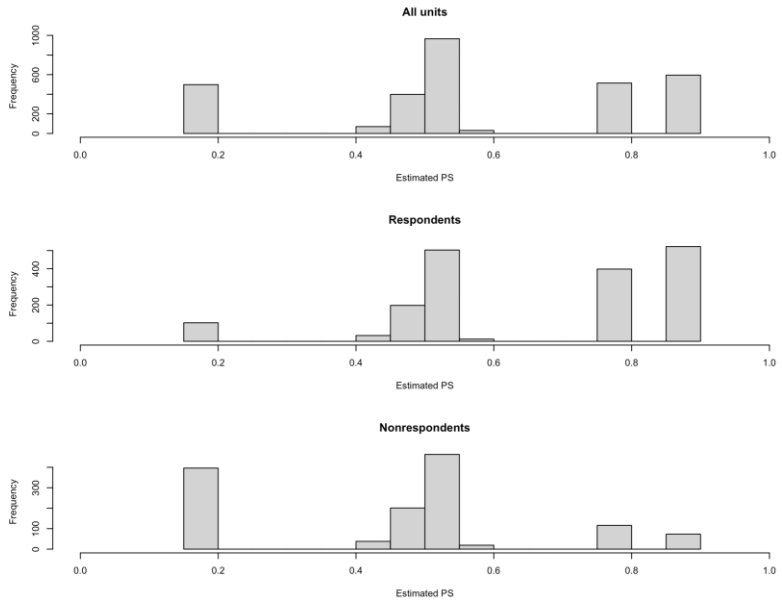
139. Another method to adjust for unit nonresponse is the propensity score method. The method uses multivariate models to estimate the response probability for each respondent, using several auxiliary variables for which values are available for both respondents and nonrespondents. The formula below shows how propensity score is estimated for the i^{th} unit.

$$\hat{\rho}_i = \frac{exp(\sum_j \hat{\alpha}_j x_{ij})}{1 + exp(\sum_j \hat{\alpha}_j x_{ij})}$$

Where x represents the value of auxiliary variables and $\hat{\alpha}$ is the fitted coefficient from the model. In this formula, logistic regression is used. Other models such as probit model, complementary Log-Log model, can also be used (Valliant et al., 2018). Now using the same two auxiliary variables in the example above, a logistic regression is fit to estimate $\hat{\alpha}$. The model coefficients are the same as in Table 5.

140. Estimated propensity score for each household in the sample, both respondent and nonrespondent, is calculated using the fitted logistic model. Bethlehem, Cobben, and Schouten (2011) claim the propensity score method can be employed if matching assumption holds. That is, for every value of the covariates, there always be respondents and nonrespondents, which yield to the following relation: $0 < \rho_i < 1$. Notice that $\rho_i = 0$ or $\rho_i = 1$ are undesirable scenarios. Figure 8 shows the histograms for the propensity score model used with Malawi's survey data. It is noticeable that for all units (both respondents and nonrespondents), there are no estimated propensity score values equal than zero or one.

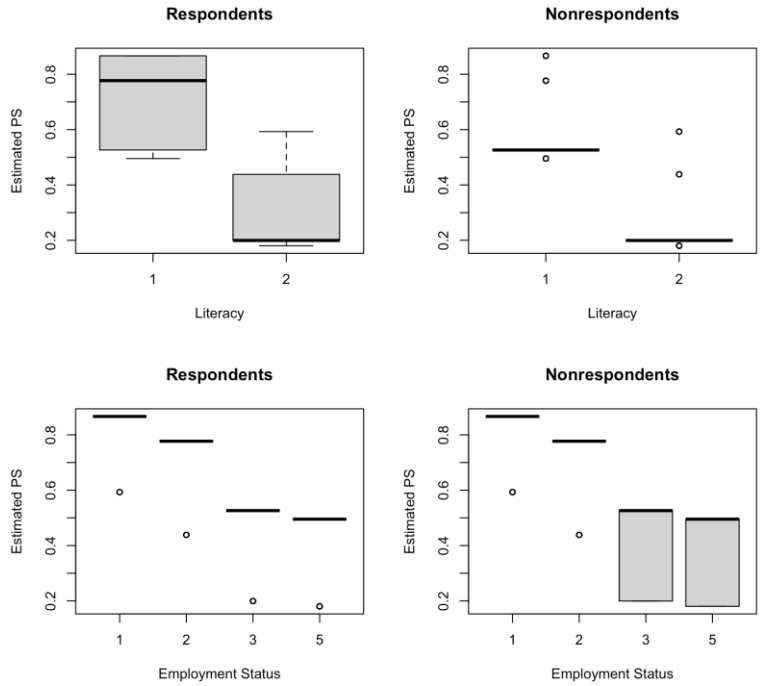
Figure 8. Histograms of the estimated propensity score for all units (upper), respondents (center) and nonrespondents(bottom).



141. Bethlehem, Cobben, and Schouten (2011) also claim that survey variables should be independent of the response behavior, conditional on the covariates of the model. Given the properties of the propensity score models this assumption can be simplified. In fact, the assumption implies that variables are independent of the response behavior, conditional on the propensity score. This property is also known as balancing. In a practical application, we can rely on this assumption if the distribution of the propensity score is the similar between respondents and nonrespondents for every single subpopulation (category) of the covariates. Figure 9 shows that the distribution of the propensity score is similar for the

two categories of the literacy covariate, and also is similar for the four considered categories of the employment status covariate.

Figure 9. Boxplots of the estimated propensity score for respondents (first column) and nonrespondents (second column) by literacy (first row) and employment status of household head (second row).



142. Classes are then created so that the response propensity for units within class is similar to each other. Therefore, the estimated $\widehat{\rho}_i$ are sorted from low to high and classes are formed with similar number of sample units within each class. In each class, one can compute different statistics, such as the mean of the unweighted response propensity, the mean of the weighted response propensity, the mean of the unweighted response rate, the mean of the weighted response rate, among others. When using these aforementioned statistics, the adjustment factors will be defined as follows, respectively:

$$a_{cps3} = \frac{\sum_{i\in\{ps\ class\}} 1}{\sum_{i\epsilon\{respondents\ in\ ps\ class\}} 1}$$

$$a_{cps4} = \frac{\sum_{i\in\{ps\ class\}} w_i}{\sum_{i\epsilon\{respondents\ in\ ps\ class\}} w_i}$$

Table 15. Four methods of estimating response propensities within classes based on fitting a logistic model, Integrated Household Panel Survey, 2019-2020, Malawi

Adjustment class	Range	Number of households	a_{cps1} Unweighted response propensity	a_{cps2} Weighted response propensity	a_{cps3} Unweighted response rate	a_{cps4} Weighted response rate
	(1)	(2)	(3)	(4)	(5)	(6)
1	[0.177,0.492]	615	0.245	0.238	0.252	0.241
2	(0.492,0.524]	614	0.509	0.509	0.489	0.479
3	(0.524,0.529]	615	0.526	0.526	0.538	0.515
4	(0.529,0.78]	614	0.731	0.723	0.728	0.692
5	(0.78,0.87]	615	0.864	0.864	0.868	0.846

Note:

- Column (1) is formed by classify households in the sample (respondents and nonrespondents) into 5 classes with the same number of households.
- Column (2) shows the number of households within each class
- Column (3) is the simple average of the estimated propensity scores within each class
- Column (4) is the weighted average of estimated propensity scores within each class, the weight used is the household base weights
- Column (5) is proportion of responding households within each class
- Column (6) is the weighted proportion of responding households within each class, the household-level base weight is used.

The function pclass in the R PracTools package is used to fit logistic, probit or c-log-log binary regressions and divide the predicted propensities into classes. More information about the function is available at https://www.rdocumentation.org/packages/PracTools/versions/1.2.5/topics/pclass; an example of using pclass is available in Valliant et al., (2018). The R code and the underlying data for this exercise is available in Annex 1.

Note: Five classes are usually used for the propensity score adjustment, but it does not have to be 5. Higher number of classes reduces biases but increases variance of the estimator. In practice, different scenarios should be tested before deciding on the number of classes to use.

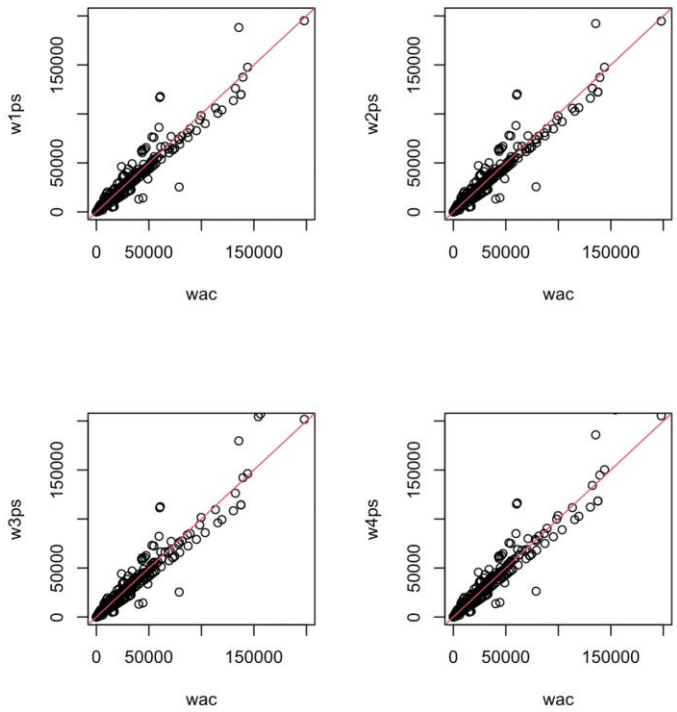
143. Table 15 above shows response probability/propensity calculation for each of the formed five classes, using four methods. One caution in forming propensity score *(ps)* adjustment classes is that many adjustment classes could potentially increase the variance of the design-based estimators. The final number of classes will depend on the mean square error of the design-based estimator. Depending on the chosen adjustment factor, the modified weight will take one of the following forms:

$$w_i^{acps\ h} = w_i^0 \times a_{cps\ h} \qquad h = 1, 2, 3, 4.$$

where $a_{cps\ h}$ corresponds to the adjustment factors calculated above.

144. For the Malawi's example, note that all methods here show similar results, so under any of the columns it would produce similar weights. This is noticeable also from Figure 10, where the relation between different propensity score adjustment is similar with respect to the previous weighting class adjustment.

Figure 10. Four different propensity score adjustments compared to the simple weighting class adjustment.



Note: The 45-degree red line represents all points where the set of weights are equal. Left upper: class adjusted propensity score by the unweighted mean ($w_i^{acps\,1}$, in x-axis) and the simple class-variable adjustment (w_i^{ac}, in y-axis); right upper: class adjusted propensity score by the weighted mean ($w_i^{acps\,2}$, in x-axis) and the simple class-variable adjustment (w_i^{ac}, in y-axis); left bottom: class adjusted by the unweighted response rate ($w_i^{acps\,3}$, in x-axis) and the simple class-variable adjustment (w_i^{ac}, in y-axis); right bottom: class adjusted by the weighted response rate ($w_i^{acps\,4}$, in x-axis) and the simple class-variable adjustment (w_i^{ac}, in y-axis).

Calibration with an external data source

145. The next step after the weighting procedures above is to calibrate the survey respondents with an external data source using auxiliary variables. This step can help correct frame coverage errors and reduce standard errors, which is particularly relevant for surveys carried out during the pandemic (see more discussion on sample frame coverage under section II.D.i). We can distinguish several cases for calibration; for example, post-stratification, raking, or generalized calibration.

146. Another important point is that for surveys that use the RDD design, there is usually not much auxiliary information available for both respondents and nonrespondents. When a telephone call is picked up by an eligible person, he or she might reject the survey interview and becomes a nonrespondent. But there is not much known about the person. Therefore, weighting adjustment covered in early part of this section might not be possible. In this case, survey practitioners might skip the previous step and move straight into the use of calibration with an external data source.

147. In the example below (*Table 16*), using the Malawian survey data, the sample from the survey was compared against 2019 Malawi demographic projections. Two auxiliary variables are used in this example – Zone of residence and Region. The post-stratification weight (column (5)) is calculated by comparing the weighted distribution of the HFPS sample by zone and region and the projection by these two variables (column (3) divided by column (4)). Note that this illustration is an over-simplified practice. As shown earlier by the time post-stratification is carried out, there should be a set of weights that have been constructed and should be applied

before the comparison with the census data. One important issue to note is that for RDD survey design, this might be the only step to benchmark the selected sample.

Table 16. Calculation of post-stratification weight, Malawi High-Frequency Phone Survey on COVID-19, 2019-2020 and Malawi population census 2018

Zone	Region	Population distribution by zone and region	Estimated population distribution by zone and region	Post-stratification weight
(1)	(2)	(3)	(4)	(5)
Urban	Northern	398,607	463,456	0.86
	Central	1,420,072	1,256,909	1.12
	Southern	1,206,235	823,443	1.46
Rural	Northern	2,267,500	1,068,992	2.12
	Central	6,504,508	2,956,848	2.19
	Southern	7,259,087	3,248,275	2.23

148. The post-stratification method is a particular case of the calibration approach, where a new set of weights w_i^{cal} is defined to comply the following restriction on specific known control totals (t_x) of the calibration covariates:

$$t_x = \sum_{i\epsilon\{eligible\ respondents\}} w_i^{cal} \times x_i$$

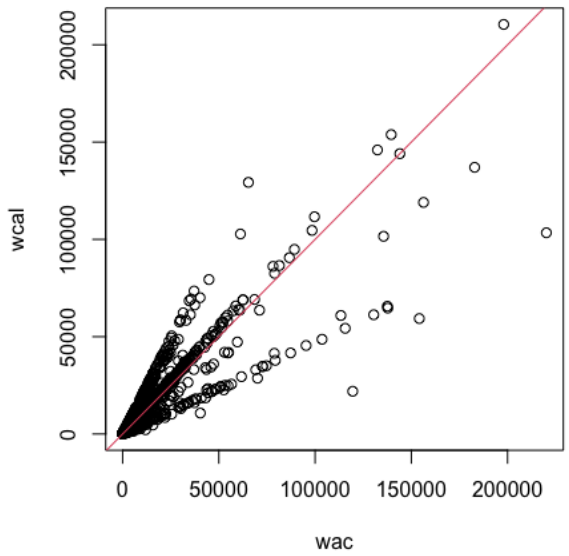
This way, the calibration weights will be defined as follows:

$$w_i^{cal} = g_i(s) \times w_i^0$$

Where, $g_i(s)$ are known as the *g*-weights. In the case of the Malawi example and considering that post-stratification is a particular form of the calibration approach, these *g*-weights are the defined in column (5) of *Table 16*.

Figure **11** shows the relation between the calibration weights and the weighting class adjusted weights. In the post-stratification case, the sum of the calibration weights will always be the population size. That is $N = \sum_{i\epsilon\{eligible\ respondents\}} w_i^{cal}$.

Figure 11. Post-stratification weighting adjustment compared to the simple weighting class adjustment.



Note: The 45-degree red line represents all points where the two sets of weights are equal. Post-stratification weights (w_i^{cal}, in y-axis) and the simple class-variable adjustment (w_i^{ac}, in x-axis).

Combined adjustment: propensity score and calibration

149. As stated by UNECLAC (2020a), coverage and selection bias are among the greatest challenges when switching the way survey data are collected. It is assumed, as a fundamental principle, that households that respond to a telephone survey do not have similar characteristics to non-respondent or uncontacted households. Changing the household survey data collection mode from face-to-face interviews to a telephone- or web-based mode may generate biases of selection, coverage, and non-response. As it was mentioned, not all the households in the original sample provided their telephone contact information; some households provided their contact information, but at the time of the interview they do not live at the selected address; some households provided their contact information, but they have since changed their contact telephone number; not all households that provided their contact information are willing to answer the survey questionnaire; among other reasons for considering that bias is far from negligible (UNECLAC, 2020b).

150. To control biases generated by considering mixed-mode surveys, the calibration approach can be applied to the propensity score adjusted weights to incorporate the best features of these two methodologies. On the one hand, a propensity score model can be used to control for covariates that were measured in a previous wave, in our example: literacy and employment status of household head. These unit-level covariates are available only for the units selected in the original face-to-face sample, and the adjustment can serve to control for selection bias and nonresponse bias in the final telephone sample. On the other hand, the calibration approach over these adjusted weights can control coverage biases while generates consistency along the estimates of the National Statistical Office.

151. Under this combined approach, a new set of weights w_i^{pscal} is defined to comply the following restriction on specific know control totals (t_x) of the calibration covariates:

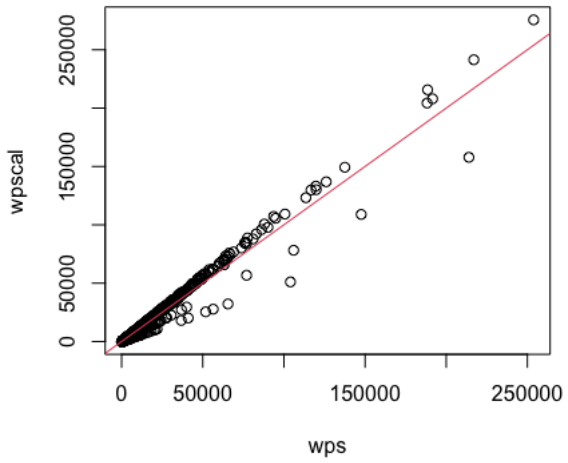
$$t_x = \sum_{i\epsilon\{eligible\ respondents\}} w_i^{pscal} \times x_i$$

This way, the calibration weights will be defined as follows:

$$w_i^{pscal} = g_i^{ps}(s) \times w_i^{acps}$$

Where, $g_i^{ps}(s)$ are the *g*-weights generated with the propensity score weights. In the case of the Malawi example, Figure 12 shows the relation between the final calibration weights and the propensity score class adjusted weights. Notice that $N = \sum_{i\epsilon\{eligible\ respondents\}} w_i^{pscal}$.

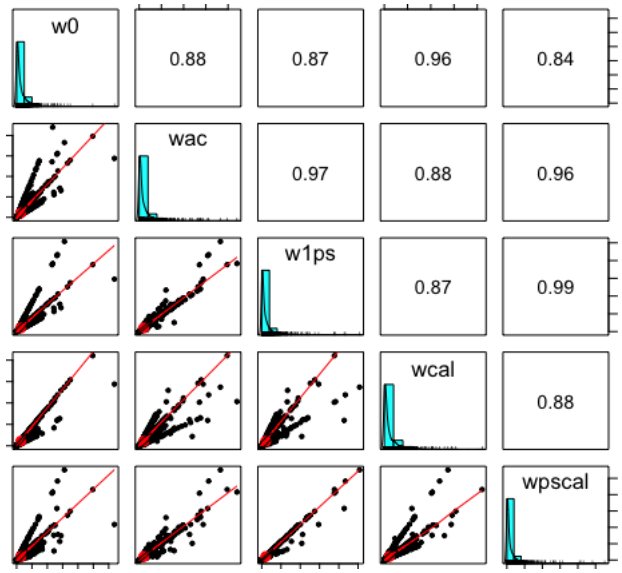
Figure 12. Post-stratification weighting adjustment compared to the simple weighting class adjustment.



Note: The 45-degree red line represents all points where the two sets of weights are equal. Calibration after propensity score adjusted weights (w_i^{pscal}, in y-axis) and the class adjusted propensity score by the unweighted mean ($w_i^{acps\,1}$, in x-axis).

152. Figure 13 shows a comparison between pairs of set of weights, ranging from original unadjusted (w_i^0), the class adjusted (w_i^{ac}), the unweighted propensity score rate in classes (w_i^{1ps}), the calibration approach (w_i^{cal}), and finally the (w_i^{pscal}) calibrated over the propensity. The diagonal of the figure shows the histogram of these set of weights, all of them have a similar distribution. Below the diagonal, the figure shows the scatterplots between the sets of weights, along with the 45-degree red line. Above the diagonal, it is shown the Pearson correlation coefficient among pairs of weights. Notice that (w_i^{pscal}) weights, which are doubly robust, are less correlated with the original unadjusted weights, and the calibration weights. That means that these two approaches by themselves are not enough to reduce the bias. However, the correlation increases with the propensity score weights, and it can be a positive signal of bias reduction.

Figure 13. A comparison of distinct set of weights. Histograms of weights are shown in the diagonal, scatterplots between pairs of weights are shown below the diagonal, and Pearson correlations between pairs of weights are show above the diagonal.



B. Mode-specific measurement effect

153. Mode-specific measurement effect refers to how the same respondent responds differently to different modes. This effect corresponds to the impact of mode changes on the measurement part of the total survey errors (Figure 2). It is usually quite challenging to measure and correct the mode-specific measurement effect unless a controlled experiment is being carried out. The entangled impact of changes in representation (mode-specific selection effect), the changes in questionnaire design and the changes in data collection mode added complications for national surveys carried out during COVID-19. In short, compared to surveys pre-COVID-19, we are looking at surveys that covered respondents who are of different profile (more affluent, living in urban areas etc.), adopted a different set of questions and a different data collection mode that may imply different responses even when the same person is responding to the same question.

154. Systematic guidance on how to tease out the mode-specific measurement effect, or the impact of changing questionnaires is not available. NSOs can review their questionnaire items and determine if any are likely to suffer from mode effects; e.g., the item was of a sensitive nature, the questionnaire wording was changed with the mode, the wording remained the same, but the ability to understand what is asked varied by mode. Schouten et al., 2021 highlighted three important methods that are useful: (a) use of paradata; (b) compare with a "gold standard" and (c) experimental design.

i. <u>Using paradata</u>

155. An increasing amount of paradata are being collected as a byproduct of the data collection process, including keystroke records, mouse-tracking, length of the interview in a telephone survey, observations of the interviewer during data collection and GPS-tracking of interviewer location. As paradata are a byproduct of a given data collection operation, the format, layout, and content of paradata are a function of the system that generated the data and may vary greatly from one form of data collection to another (Kreuter, F, 2013). As explained in Kreuter, F. (2013), paradata can be used to provide indications of how individuals responded to questions and the processes invloved. For example, a long response time could be an indication of a lack of

knowledge of respondents but also a complex question design and short response time could imply a good knowledge of respondents and easy-to-understand questions but could also be associated with interviewer falsifications or "respondent speeders". The challenge with using paradata to compare two data collection modes, one of them being face-to-face, is that the availability of paradata is different across the mode. Face-to-face interviewing using PAPI does not collect as much paradata as other collection modes.

156. For countries that switched from face-to-face interviewing using CAPI to CATI during COVID-19 CAPI is often still used to capture data by the interviewer. In this case the same types of paradata are being captured through the CAPI platform. It is then possible to compare the paradata and see whether there is any significant difference for either the same respondent or respondents of similar characteristics. There is unfortunately very little available in the literature to demonstrate how this could be done. When Statistics Austria tested the use of the web for its LFS, paradata used to monitor the quality included response time, pauses, help menu access, among others. The results show that certain questions had a high proportion of help menu accesses than others. (Hartleib et al., 2021)

ii. Comparing with pre-COVID-19 results

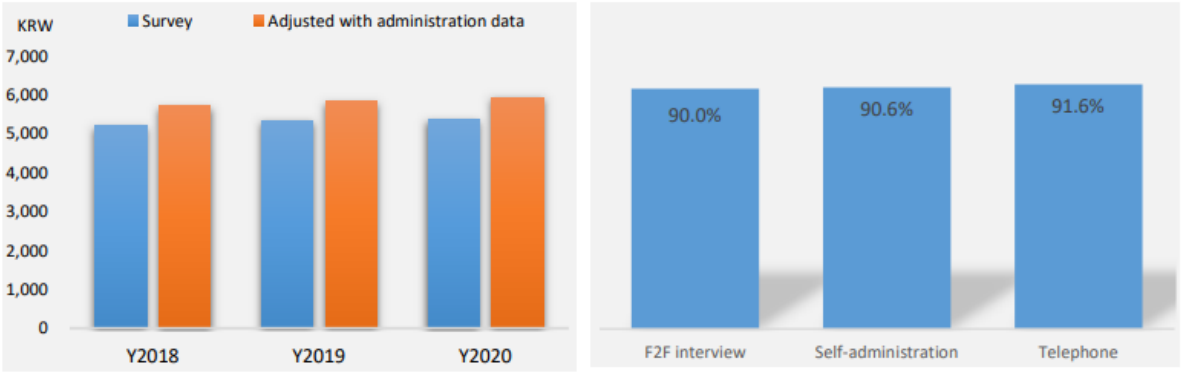
157. If there exists information for a closely related item, there can be the potential to adjust using it as a "gold standard" reference. Certainly, there is no real gold standard, but in this case, we are comparing estimates with those that were compiled pre-pandemic. This might be available from a reliable external source of information, either an administrative source or other surveys, possibly at an aggregated level (totals for a village, which can be compared against survey totals).

158. The annual Household Finances and Living Conditions Survey (HFLC) carried out by Statistics Korea (KOSTAT) collaborates with administrative agencies in charge of tax, health insurance and social insurance to adjust data produced by HFLC (Kim et al., 2021). During the pandemic, HFLC experienced a large shift in survey mode: from only 4 per cent in 2019 responding with non-contact survey mode (telephone, internet, fax or self-administered paper-based) to 22 per cent in 2020. To assess the effect of the changing mode on data collected in HFLC, the level of adjustment required using administrative data on income which was compared for 3 years from 2018 to 2020 (Figure 14). Panel *a* on the left shows a similar pattern of adjustment required from 2018 to 2020, suggesting a non-significant mode impact. Panel *b* shows the income coverage rate[13] by survey mode. Telephone mode provides a better estimate than face-to-face interviewing, which could be explained by the difference in households responding to those modes. Households using the telephone mode have higher education levels and higher paying jobs than those who responded using the face-to-face mode. This again shows that mode effect is not significant.

[13] Coverage rate is calculated as the ratio of survey estimate of income to the final adjusted income level after administrative data are integrated. Higher coverage rate implies better estimates from the survey.

Figure 14. Adjustment of household income using administrative data, by year and by mode of data collection, Household Finances and Living Conditions Survey, Republic of Korea

a. Income from survey results over three years b. Income coverage rate by survey mode



Note: Figure 5 in Kim et al. (2021).

iii. Designing an experiment

159. Those NSOs with the resources should consider a short-term bridge study where they conduct the study in parallel, using both the old and new methodology for the same period. "One sure way to separate the change in collection mode from the effect of the pandemic is to maintain telephone and mixed-mode surveys in the post-pandemic period for a reasonable time." (ECLAC 2020a) Bridge studies are a methodology often used for long time series when an updated sample or other methodological change is introduced; for example, when updating the sample and methodology used to estimate unemployment rates (Norwood and Tanur, 1994) or fishing and hunting statistics (Erciulescu, et al., 2021). Such studies "are particularly appropriate if the sample data observed under the regular and the new approach are not consistent" (van den Brakel et al., 2008). That is exactly the situation with the pandemic, where some of the underlying characteristics may be different than they were before the pandemic. Van den Brakel et al. (2008) discuss alternative designs for such studies, which depend on a number of factors including the ability of interviewers to conduct the study using both methodologies. If only interested in changes to the questionnaire, a bridge study could be incorporated into the existing face-to-face data collection by randomly using the shortened or regular questionnaire. A bridge study is costly but could provide very useful insight into the comparability of the estimates from the different data collection methods.

160. To conduct a bridge study, one should follow these basic steps:

- For a specified data collection period (e.g., one year, one quarter) conduct the data collection using both the pandemic and standard methodology. This may mean face-to-face and telephone surveys at the same time.
- For this period, select a larger than normal sample because only half will receive the "standard" methodology. Assign half of the sample to each. Ideally, half of each geographic area will receive each method, but the NSO will have to decide on the smallest geographic level that can be assigned to either (e.g., village, enumeration area).
- Compute estimates separately from the two methods and compare them. The differences are generally assumed to be due to methodological changes, not the pandemic. These can be used to adjust data from either series to achieve a more consistent time series and avoid, bridge or at least reduce breaks in series.

161. Re-interviews (Schouten et al., 2021) are an alternative method for estimating the impact of mode-specific measurement effects. Weighting adjustments minimize bias by correcting to known totals from administrative or other large data sources. But these known totals may not be highly correlated with items that are subject to measurement effects. If in a survey there are items that are correlated with the likelihood of having mode effects, then re-interviews can leverage those to improve adjustments beyond what is possible by weighting.[14]

- Once face-to-face interviewing has returned, re-interview a random sample of those who participated in the pandemic data collection, using the standard data collection method. Although this has been commonly used to assess the mode effect, caution needs to be exercised when comparing data collected during and after COVID-19. For the same sample, the estimate of key variables could change due to the pandemic or memory lapse rather than the mode effect.
- Assuming the standard data collection (face-to-face) is the closest to the true values, we can use the responses to other items, along with sampling frame information to adjust everyone's telephone responses to be more consistent with face-to-face, not just those that are re-interviewed.
- Re-interviewing assumes that the relationship between the face-to-face response for household and the telephone response can be described as a function of covariates available from the survey or frame.
- The estimated total (that would have been achieved by face-to-face interviewing) from telephone data collection during the pandemic is then estimated using the relationship established in the above step.

162. The Netherlands traditionally collected their LFS using face-to-face, telephone, and web data collection. When the pandemic happened, they lost the ability to use the face-to-face mode, but their multi-mode history of the survey allowed them to model how much of the change in estimates across time was due to this mode switch (Van den Brakel et al., 2021). Using the mode-effect evidence pre-COVID-19 to adjust for the loss of the face-to-face mode, the Netherlands were able to show that the underlying changes in the economy due to the pandemic were the main cause of changes to estimates.

163. Mode-specific measurement effect has also been tested without a controlled experiment. To assess the difference in web and telephone mode for selected variables in the LFS, the National Institute of Statistics and Economic Studies of the Grand Duchy of Luxembourg (STATEC) carried out a study that approximated a randomized experiment using the Coarsened Exact Matching method (Iacus et al., 2012) with the collected data from 2015 to 2017 (Schork et al., 2021). The study found the difference in reported employment status between web and telephone respondents was due to the differences in coverage and nonresponse under the two data collection modes (i.e., selection bias). There is no measurement bias for this particular outcome indicator. However, subjective indicators such as wage adequacy and job satisfaction seem to be more impacted by the mode-specific measurement effect.

C. Major effects outside of the impact of changes in data collection mode

164. So far, the chapter has only covered quality assessment and possible corrections due to the change of data collection mode (Section III.A on mode-specific selection effect and Section III.B on mode-specific measurement effect). Other changes occurred during the pandemic, as

[14] Re-interivewing households for mode effect needs cautious

discussed in Chapter II, in terms of changes in the survey questionnaire, training, supervision, data collection protocols, data entry and quality assurance, have not been covered. These aspects will have an impact on the data quality, but it is difficult to assess the magnitude of the impact and what was the underlying cause and which one contributed the most in terms of quality. If we were able to compare data collected during COVID-19 with a good-quality survey pre-pandemic, the difference shown in the estimate would be the result of all changes introduced to the survey during the pandemic plus the real change that occurred. According to Statistics Canada, subject areas (covered by its 2021 population census) that are likely to be impacted by the pandemic include labour and commuting, income, housing, immigration, family, households and marital status (Statistics Canada, 2021). This means that even without any changes in survey instruments and how data were collected during COVID-19, the collected data for these subject areas would have been considerably different from those collected during the non-pandemic period.

165. There are also quality dimensions falling outside the TSE framework that need to be considered when evaluating surveys. Within the UN-NQAF, principles of timeliness and punctuality, and coherence and comparability might have been particularly affected during the pandemic.B

166. Given that many NSOs had to cancel or delay data collection (half of the countries conducting an LFS (Discenza and Walsh, 2021) and many other countries delayed less established data collections), the timeliness of reporting results was frequently affected by the pandemic. Sixty-five percent of countries reported that their publications and data dissemination were delayed because of the pandemic (United Nations/World Bank, 2020b). Similar findings were reported for countries in Latin America and the Caribbean, where 60 percent of the survey data dissemination was delayed (UNECLAC, 2022).

167. There were, however, a few examples, where timeliness was improved during the pandemic. Malaysia (Ali, 2021) expedited the dissemination of their Monthly Labour Force Report from 6 weeks down to 5 weeks. Grenada (Brizan, 2021) also improved timeliness in response to the urgent need for data by the government.

168. With respect to comparability, the pandemic created an abrupt change in both data collection methodology and true conditions in countries around the globe. This created serious issues in comparing data across time. As mentioned earlier, almost every NSO stopped face-to-face data collection in May of 2020, but by October of same year, a quarter had fully and another 40 percent had partially resumed face-to-face interviews. This introduces a wide range of potential sources of error that are unique to this time period. This does not automatically make these data noncomparable to earlier and later estimates, but it does require careful documentation so that users are aware of these issues and what has been done to mitigate them.

169. Some countries cancelled data collections completely for multiple months. Repeated surveys (including panel surveys such as the LFS) typically collect data spread across the year; for example, quarterly surveys are designed to capture seasonal differences. If countries report annual data for 2020 where a set of months are missing, this can introduce biases if those months have atypical conditions (e.g., lower employment rates). By excluding those months, the reported annual statistics will not truly represent the entire year, introducing insufficient comparability with other years.

170. The time period during a month over which the information was gathered, or the reference period for specific items, may have changed as well (Discenza and Walsh, 2021). These can also lead to comparability issues that need to be understood by users. Comparing these reference periods to the timing of COVID-19-related events can provide useful insight into any comparability concerns.

D. Summary

171. The chapter discusses ways to assess possible impact of changes introduced on household surveys during the pandemic. A large proportion of the chapter covers the assessment and reduction of the mode-specific selection effect that stemmed from sample frame deficiencies and increased nonresponses due to the change of data collection mode. Assessing mode-specific measurement error is not an easy task, typically requires an experiment to fully tease out the potential impact. The chapter then covers possible ways to detect mode-specific measurement errors through the use of paradata, comparing with gold standard or designing an experiment.

172. Data coherence and comparability are also covered briefly in the chapter. Other errors introduced due to changes in survey questionnaire, or other changes covered in Chapter II, section D.iv, require specific assessments. There are some discussions on assuring quality in Chapter II. For example, some countries carried out random checks of completed telephone interviews to ensure that all interviews were actually carried out. But errors due to different changes introduced in survey operations tend to be confounded, and controlled experiments are usually required to fully tease out various impacts.

173. For each challenge caused by the shift of survey mode during the pandemic and addressed in the chapter, Table 17 summarizes the assessment steps to follow and the approaches proposed for each effect. Following the principle that there is no "one-size-fits-all" approach, NSOs might test and choose the most appropriate approach.

Table 17. Summary table on methods for quality assessment and error correction

Effect of shifting mode	Assessment steps	Possible assessment and correction methods
Mode-specific selection effect	Detecting bias	*Bivariate test of independence (Pearson's Chi-squared Goodness of Fit, Cramer's V test)*
		Visualization tools
		Multivariate analysis (logistic regression)
		Comparing sample and population
		Representativity indicator
		Disaggregated analysis
	Select benchmarking source (e.g., panel, cross-sectional)	*Censuses (coverage and measurement errors)*
		Administrative sources (coverage, measurement and sampling errors)
		Probabilistic surveys
	Correct bias	*Weighting class adjustment*
		Propensity score adjustment
		Calibration with external data sources
		Combined adjustment
Mode-specific measurement effect	Requires controlled experiments (no specific guidelines)	*Paradata*
		Gold standard comparison
		Design experiment (bridge study, self-report, re-interview)

Non-mode related errors	*Not covered*
Timeliness and punctuality	*Briefly*
Coherence and comparability	*Briefly*

IV. Surveys based on non-probabilistic sampling

174. The entire set of sampling theory and error estimation is based on a probability-sampling framework. Within this framework, each sample has a known selection probability and units have an associated inclusion probability, such that inference can be made accordingly. During COVID-19, a large number of surveys suffered from large and outstanding nonresponse rates. With this substantial quality decrease, it is difficult to assume that the final subset of respondents still reflects a representative probabilistic sample. At the same time, many surveys were carried out using non-probability sampling. This choice often stemmed from necessity when there was no frame readily available for sample selection, but there was a pressing demand for timely data.

175. At a time that there has been a lot of debate over the trade-offs between quality and timeliness, a discussion would be useful to discuss what non-probabilistic surveys could and could not do; and what needs to be considered when using a non-probabilistic survey. First of all, not all non-probabilistic surveys are created equal. As noted by AAPOR (2013), different non-probability sample approaches can be seen as falling on a "*continuum of the expected accuracy of the estimates*". Therefore, it is difficult to have a unified response to the use of nonprobability samples: assessment needs to be made in a case-by-case manner.

176. Another important element when discussing the usefulness and validity of non-probabilistic survey is the objectives of the survey. If it is a survey that is with a relatively narrow focus (e.g., voting preferences), then the results could be useful with a good set of covariates for benchmarking. However, for surveys that collect a large number of outcome indicators and aim to meet multiple objectives, this would require a large number of covariates for benchmarkin which are often not available.

177. Statistics Canada carries out crowdsourcing web data collection, to obtain citizens' input on their priorities; to collect data in emergency situations such as COVID-19; and to gather information when there is no other source available (Statistics Canada. 2022b). The data collection is nonprobablistic yet has provided valuable information during the pandemic.

178. The following aspects about the use of non-probability sampling are extracted from AAPOR (2013). Readers are encouraged to explore more information from the original source.

- *Transparency of the methodology used is essential*. It is extremely important for survey managers to be transparent and communicate the method used to select the sample, an assessment of the "frame" (e.g., telecommunication company and the coverage of its services), the coverage issue with the frame and strategies to resolve them; e.g., what is the percentage of mobile phone ownership or internet access in the country? Does telephone ownership vary by region or population groups? What methods are used to produce the final estimates? What data sources are used to correct the potential biases?
- *The validity of non-probability samples for statistical inferences depends greatly on the underlying model assumptions*. As noted in the AAPOR (2013) study, the missing link (the selection probability) between the sample and the population it represents is a complicating factor to make inferences about the population using the non-probability sample. In the recent study carried out by Pew Research Center (2018),

choosing the right variables for weighting/calibration is the key in reducing bias from an online opt-in sample survey. Methods used to correct the bias, whether simple or more sophisticated, do not affect much the quality of the final estimates.

179. In order to get appropriate inferences from nonprobability samples, Valliant (2020) describes four approaches that heavily rely on auxiliary information. These are described as follows:

- *Quasi randomization*. This approach relies on the fact that a recent probability sampling (reference sample) is available, and the main objective of this method is modeling pseudo-inclusion probabilities and corresponding weights in the nonprobability sample that mimic the performance of the probabilistic sample. Unlike in probability sampling, the selection mechanism of units is unknown, and pseudo-inclusion probabilities of being in the sample are estimated by using a working model that relates the inclusion in the nonprobability sample with a set of known covariates in both, the reference and the nonprobability sample.

  Assuming that the working model induces estimates of the pseudo-inclusion probabilities $\widehat{\pi}_i$ in the nonprobability sample s, that depends on the covariates X_i and the response mechanism, such that $\widehat{\pi}_i = Pr(i \in s \mid X_i, \phi_i)$, then the new set of weights are given by the inverse of the estimated pseudo-inclusion probabilities. That is: $w_i^{qr} = 1/\widehat{\pi}_i$.
- *Superpopulation modelling*. This completely model-based approach defines the observed values of the outcome variables as resulting from a superpopulation model, which is estimated directly from the nonprobability sample. For example, a linear model can be written as follows:

  $$E_\xi(Y_i) = X_i'\beta \quad \text{and} \quad V_\xi(Y_i) = v_i$$

  Once the model is fitted, the next step is to predict the values of the variable of interest in the units that were not included in the sample. This way, the prediction of the variable of interest in the set of non-sampled units will be given by $\widehat{E}_\xi(Y_i) = f\left(X_i'\widehat{\beta}\right)$. A predictor of the population total, is given by the following expression:

  $$\hat{t}_y = \sum_{i \in s} Y_i + \sum_{i \notin s} Y_i = \sum_{i \in s} w_i^{spm} Y_i$$

  Where, $w_i^{spm} = 1 + \left(t_{x_u} - t_{x_s}\right)' \left(X'X\right)^{-1} x_i$, and t_{x_u} denotes known control totals of the covariates used in the model; in the same way, t_{x_s} denotes the sample total covariates involved in the model. Note that this set of weights w_i^{spm} is unique in the sample. This way, inferences can be carefully done if the model holds for the outcomes of interest. Also notice that control totals should be known for this method to work. The more the covariates relate with the outcome of interest, the better the prediction.
- *Doubly robust estimation.* As noted by Valliant (2020), the aim of this approach is to define approximately unbiased and consistent estimators when the pseudo-inclusion distribution, the superpopulation model, or both are correctly specified. First, pseudo-inclusion probabilities and corresponding weights w_i^{qr} are estimated as described above; then they are used in conjunction with a superpopulation model to compute doubly robust weights. For instance, in the case of a linear model, the final weights take the following form:

$$w_i^{dr} = w_i^{qr}\left(1 + \left(t_{x_u} - \hat{t}_x\right)' \left(X'WV^{-1}X\right)^{-1} X_i/v_i\right)$$

Where, W, and V are diagonal matrices containing the sample quasi randomization weights, and the sample estimated variances of the superpopulation model, respectively. Notice that $\hat{t}_x$ is the weighted estimate for the total of the auxiliary variables.

- *Multilevel regression and post-stratification (MRP)*. As claimed by UNECLAC (2020a), this is a useful technique for predicting a parameter of interest within small domains by modelling the mean of the variable of interest conditional on poststratification counts. The MRP model is composed of two parts: the first involves adjusting a multilevel regression model based on the nonprobability sample; and the second is the poststratification stage that uses census counts to adjust for selection bias. From Gelman (2007, eq 7) it can be shown that the resulting MRP weights are giving by the following expression:

$$w_i^{mrp} = \left(n(N^{pop})' \; X^{pop}\left(X'X\right)^{-1} X_i/v_i\right)$$

Where, $N^{pop} = (N_1, \ldots, N_J)$ represents the census counts of the *J* post-stratification cells, and X^{pop} is the matrix of predictors for the *J* poststratification cells.

V. Disseminating Survey Data During COVID-19 and Communicating Data Quality

A. When and what to publish?

180. Depending on the quality of data, countries may decide whether the data would be published and determine the amount of the data to publish. UNECLAC (2020c) suggested three scenarios for countries to consider when deciding to publish the indicators or not:

a) Poor levels of coverage, for which it was recommended not to use the data collected for the production of any official statistics.

b) Average levels of coverage, for which it was recommended to modify the expansion factors in the survey to limit the levels of bias and publish official statistics only at the national level.

c) Acceptable levels of coverage: where it was recommended to continue with the normal publication process.

181. Countries that do not meet the minimum sample size ~~(e.g., 1,800 households)~~ for national-level estimate are discouraged from publishing any data from the survey (case a). If sample size is larger than the threshold for national estimate, countries may consider publishing data only at the national level, while adopting procedures to correct potential biases (case b). Imputation was discouraged even though information may be available to do so. The guidance also recommended proper communication on the differences in data quality when data are published.

B. Communicating survey methods and results

182. Communicating survey methodology and how data should be used is always an integral part of the survey data dissemination stage. Accurately documenting the data collection process, then disseminating information about the quality impacts of COVID-19 on the data will be vital in retaining the NSOS's credibility with users (Credibility is included as one component by the OECD quality framework (OECD, 2011)). For surveys carried out during COVID-19, additional information needs to be provided to users, including:

- Changes introduced in the survey including how the questionnaire has changed, the change in data collection mode, coverage of the frame and response rate
- How these changes are likely to impact survey data quality
- Quality control mechanisms used before, during, and after the pandemic that impact quality
- Methods used to minimize the impact of COVID-19 and changes on data quality
- How users should interpret and use the data. Note that during this time period changes in survey results could be a mixture of real social and economic effects due to the pandemic and methodological impacts on data collection. It is important to communicate this to the user community.

183. Describing these will communicate both the strengths and weaknesses, in the data series. The Dutch paper mentioned above (van den Brakel, 2021) is such a discussion, although it aimes at more technical users. Counties might choose to include a special methodological section when reporting pandemic data that explains what are unique to these data and how best to interpret results. This is an opportunity for data users to better understand the complications that are regularly confronted, and the need for ongoing support of the NSO as it strives to continuously improve in the future.

184. The examples below showed how countries communicated the survey results with the public. In Box 2, information is provided on (a) time the survey was carried out; (b) mode of data collection; (c) how sample was selected (from the 2010 census); (d) target population (15 years and older, in private homes); (e) reference period used for labour force data collection; (f) level of representation (national); and (g) warning about comparability with pre-COVID-19 data series. The National Institute of Statistics and Census in Panama also publish data on sampling errors for key labour force indicators (INEC - Panama, 2020b). The Uruguay National Institute of Statistics (INE) also noted in its data dissemination that until studies are carried out to rule out or measure the biases produced by changes in the operation of the survey, the estimates offered by the non-face-to-face are not strictly comparable with the face-to-face Permanent Continuous Household

Box 2. Communicating difference in the LFS, September 2020, Panama

The National Institute of Statistics and Census (INEC) carried out in September 2020, the Telephone Labor Market Survey (EMLT), in the midst of the health crisis that the whole world is experiencing, and Panama does not escape from it. Due to the biosecurity restrictions and standards imposed by the health authorities, this year, the data collection was carried out, through telephone calls to informants. The preparation of the sample and the final population estimates are based on the 2010 Population and Housing Censuses.

The study universe of the survey is the population of 15 years of age and older, who usually reside in private homes. The data obtained have as a reference, the week that precedes the one in which the interviews are carried out. Consequently, the figures correspond to a weekly average of the preceding months. During the months of September-October, telephone calls were made to households for 95% coverage.

Due to the methodological changes in the collection of survey data and the adjustments to the sample in the midst of the health crisis, information is available for the characterization of the Panamanian labor market, among other aspects, at the national level. The series of the Labor Market Survey prior to 2020, can only be taken as referential elements, since in the face of the unprecedented health and economic crisis, there is a rupture of the statistical series for the measurement of the labor market.

The information provided is representative with sample sizes at the national level. The data of the Telephone Labor Market Survey 2020 are published through the website of the Comptroller General of the Republic.

This office expresses its gratitude to the members of the households selected in the telephone sample, for the cooperation offered in the provision of the data, for the elaboration of these statistics.

Source: INEC - Panamá (2020a).

Survey (ECLAC, 2022).

185. The Household Pulse Survey carried out by the US Census Bureau is a 20-minutes online survey studying how COVID-19 is impacting households across the country. Since its first phase implemented in April 2020, 7 phases of the survey have been completed. For each one of them, technical documentation is published that includes methods, accuracy, user note and questionnaire (United States Census Bureau, 2022). Box 3 provides an example on how the US Census Bureau communicates the quality of the 2020 Household Pulse Survey to the public. Comparability of data noted that "*Data obtained from the HPS and other sources are not entirely comparable. This is due to differences in data collection processes within this survey and others. These differences are examples of nonsampling variability not reflected in the standard errors. Therefore, caution should be used when comparing results from different sources.".*

Box 3. What is included in the Source of Data and Accuracy of the Estimates for the 2020 Household Pulse Survey

Source of data: explains the objective of the survey, collaborators, frequency, mode of data collection
Sample design: frame, sample size, sample rotation, obtaining contact information, availability of contact information for frame elements, data collection platform, response rate
Estimation procedure: how weights are created including adjustment for nonresponse and undercoverage, personal adjustment and raking to population estimates
Accuracy of estimates: sampling and nonsampling error, nonresponse, undercoverage and standard error calculation. Comparability of data and a warning on nonsampling error are discussed.
Source: United States Census Bureau, 2022

VI. Lessons learnt and implication for future

186. Sudden changes forced by the outbreak of the COVID-19 pandemic certainly affected most of the countries in various ways making the impact of such extraordinary event difficult to assess. This Guidance Note draws from literature and recent studies a set of tools that national statistical offices might consider and adapt to their case. In some cases, the presence of bias, coverage or sampling errors might be evident, for example if the mode, the questionnaire and the period of the survey have changed. In other cases, for example in countries where CATI was already in place, the effects of the pandemic on surveys might be driven by less evident causes, e.g., psychological reasons, sudden relocation of individuals or households.

187. In Chapter **Error! Reference source not found.**, we provided an overview of the potential changes introduced in household surveys during the pandemic, which was largely drawn from the experiences of national statistical offices. These changes were then linked to the classical total survey error framework (TSE) as well as, to a lesser extent, the UN National Quality Assurance Frameworks for Official Statistics to provide a broad picture on how different element of errors are to be impacted by the changes introduced during the pandemic. We hope readers of this Guidance Note could relate and position themselves in this discussion and find it helpful.

188. Chapter **Error! Reference source not found.** presents examples and approaches to assess and reduce biases and other mode-specific effects. Sometimes, like in the case of mode-specific measurements effects, it is difficult to make ex-post assessments. But, in other cases, it could still be possible to carry out controlled experiments while getting back to a new household survey framework. With respect to the mode-specific selection effect, this part provides the highest number of suggestions and examples that NSOs could implement to detect and assess biases and errors. Running different tests and analysis on surveys and gathering results from

different countries could be a precious source of information to assess the impact of the COVID-19 pandemic and provide material for further research and progress.

189. Surveys based on non-probabilitic sampling (Chapter IV, sometimes referred to as crowdsourcing) have been used more frequently during the pandemic. The choice for non-probabilistic sampling could be out of necessity as there is no existing sampling frame to contact the respondent. But even for countries that do have the necessary survey infrastructure, nonprobabilistic surveys were also conducted to fill a data gap that the traditional survey is not able to meet the user needs (Statistics Canada, 2022b). The chapter covers what countries should consider especially in terms of communication when carrying out surveys based on non-probabilistic sampling and offers ways to benchmark the collected data.

190. Chapter V is on communication and dissemination – which is considered extremely important as a step to inform users about the data product. It is crucial that NSOs convey in a clear way to the public, and especially to current and potential participants, their results, challenges and limitations. This would enhance the household survey effectiveness "transforming respondents into collaborators and co-producers" (ISWGHS, 2022b). Transparency remains crucial when communicating with the users.

191. In summary, the Guidance Note provides, through concrete examples and general principles, a set of tools that might guide NSOs in (a) assessing the quality of surveys carried out during the pandemic; (b) reducing or minimizing, as much as possible, the effects of changes in the data collection mode; (c) communicating and disseminating the results. We certainly believe the Guidance Note would be useful for countries interested in having a good understanding on the quality of surveys carried out during the pandemic. But it also becomes obvious that ex-post assessment and adjustment have serious limitations. For example, mode-specific measurement effects are very difficult to assess without controlled experiments. The impact of other changes such as training of enumerators, quality assurance procedures or most of the changes covered in Section II.C.iv cannot be assessed once these changes occurred. Anecdote from countries is helpful to gain insights on what has happened during the pandemic but assessing its impact on data quality is challenging if not impossible, unless a controlled experiment is being carried out.

192. Therefore, we also hope this Guidance Note can serve as a wake-up call for countries to use the opportunity to be more prepared for the next crisis. For example, implementing bridge studies to assess the changes from the methods used during the pandemic to the post-pandemic methods; if moving towards a mixed-mode survey, designing the switch properly with proper testing; building or maintaining an updated contact list in a coherent way. All these are examples of how pandemic could actually become an opportunity for the future of surveys. As clearly laid out in the paper prepared by the Inter-Secretariat Working Group on Household Surveys on "Positioning Household Surveys for the Next Decade", "*strengthening NSO capacity in remote data collection specifically in low- and middle-income countries is a key strategic step to ensure that phone and web surveys can be used together with their face-to-face counterparts, both to rapidly respond to data needs in the aftermath of shocks or to increase the frequency and timeliness of survey data collection during emergencies*." (ISWGHS, 2022b)

VII. References

Ali, N. M. (2021). Malaysia's labour market information: accelerating improvement amidst COVID-19 pandemic. Asia-Pacific Stats Café.

Ambel, A., McGee, K., and Tsegay, A. (2021). Reducing bias in phone survey samples: effectiveness of reweighting techniques using face-to-face surveys as frames in four African countries. World Bank Policy Research Working Paper 9676. https://documents1.worldbank.org/curated/en/859261622035611710/pdf/Reducing-Bias-in-Phone-Survey-Samples-Effectiveness-of-Reweighting-Techniques-Using-Face-to-Face-Surveys-as-Frames-in-Four-African-Countries.pdf

American Association for Public Opinion Research (AAPOR, 2010). Cell Phone Task Force Report: new considerations for survey researchers when planning and conducting RDD telephone surveys in the US with Respondents reached via cell phone numbers. Available at https://www.aapor.org/Education-Resources/Reports/Cell-Phone-Task-Force-Report.aspx

American Association for Public Opinion Research (AAPOR, 2013). Non-probability sampling: report of the AAPOR Task Force on Non-Probability Sampling. Available at https://www.aapor.org/Education-Resources/Reports/Non-Probability-Sampling.aspx

American Association for Public Opinion Research (AAPOR, 2016). Address-based Sampling. Available at https://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx#1.2%20What%20is%20Addressbased%20Sampling?

American Association for Public Opinion Research (2016). Standard definitions: final dispositions of case codes and outcome rates for surveys. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: a field experiment. Public Opinion Quarterly, 58, 210-240.

Beaumont, J. and Rao, J.N.K (2021). Pitfalls of making inferences from non-probability samples: can data integration through probabilistic survey provide remedies? The Survey Statistician, 2021, Vol 83, 1-22.

Bethlehem, J., Cobben, F., & Schouten, B. (2009). Indicators for the Representativeness of Survey Response. Statistics Canada's International Symposium, 10.

Bethlehem, J., Cobben, F. and Schouten, B. (2011) Handbook of Nonresponse in Household Surveys. Volume 568 of Wiley Handbooks in Survey Methodology. John Wiley & Sons, 2011.

Biswas, S. (2020). Coronavirus: India's pandemic lockdown turns into a human tragedy. British Broadcasting Corporation. https://www.bbc.com/news/world-asia-india-52086274

Blumberg, S. (2021). Impact of the Pandemic on National Health Insurance Survey Data Collection. 2021 FCSM Research and Policy Conference.

Brizan, H. (2021). Discussion with the Grenadian statistical office.

Brooks, C.A. and Bailar, B.A. (1978). An error profile: employment as measured by the Current Population Survey. Washington D.C: U.S. Office of Management and Budget (Statistical Policy Working Paper Number 3).

Brubaker, J., Kilic, T. and Wollburg, P. (2021). Representativeness of individual-level data in COVID-19 phone surveys: findings from sub-Saharan Africa. https://doi.org/10.1371/journal.pone.0258877

Carletto, G. (2020). Survey data collection during and after COVID: lessons and recommendations. Presentation at the Joint ISWGHS-ECLAC webinar on COVID-19: assessment of the mode effect on official statistics, December 2020.

Carletto, G., Chen, H., Kilic, T., and Perucci, F. (2022) Positioning household surveys for the Next Decade. Statistical Journal of the IAOS – 1 (2022) 1-24.

Dillman, D. A., and Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. Field Methods, 17, 30-52.

Discenza, A.R. and Walsh, K. (2021). Lessons from the COVID-19 Pandemic: Global review of impacts of the COVID-19 pandemic on labour force surveys and dissemination of labour market statistics. Department of Statistics, International Labour Organization. https://ilo.org/wcmsp5/groups/public/---dgreports/--stat/documents/publication/wcms_821387.pdf

East, S., King, W., Payne, C., Vassilev, G., and Wallace, S. (2021). The ONS online time use survey: creating a feasible platform for 21st century time use data collection. Paper prepared for the 26th IARIW Virtual General Conference, August 23-27, 2021. Available at https://iariw.org/wp-content/uploads/2021/08/TheONSOnlineTimeUseSurvey_paper.pdf

Edwards, B. (2021). Personal communication about procedures used at Westat.

Ehling, M. and Körner, T. (2007). Handbook on Data Quality Assessment Methods and Tools. Eurostat. Available at: https://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf

Erciulescu, A.L., Opsomer, J.D., and Breidt, F.J. (2021). A bridging model to reconcile statistics based on data from multiple surveys. *Annals of Applied Statistics* 15:2, pp. 1068-1079. DOI: 10.1214/20-AOAS1437. https://projecteuclid.org/journals/annals-of-applied-statistics/volume-15/issue-2/A-bridging-model-to-reconcile-statistics-based-on-data-from/10.1214/20-AOAS1437.short

Eurostat (2007). Handbook on Data Quality Assessment Methods and Tools https://ec.europa.eu/eurostat/documents/64157/4373903/05-Handbook-on-data-quality-assessment-methods-and-tools.pdf/c8bbb146-4d59-4a69-b7c4-218c43952214

Frost, R. (2021). Have more people moved during the pandemic? *Joint Centre for Housing Studies of Harvard University*. Available at: https://www.jchs.harvard.edu/blog/have-more-people-moved-during-pandemic#:~:text=There%20were%2031.74%20million%20permanent in%20both%202019%20and%202020

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. Statistical Science, 22(2). https://doi.org/10.1214/088342306000000691

Ghana Statistical Service (GSS), Ghana Health Service (GHS), and ICF International. 2015.

Ghana Demographic and Health Survey 2014. Rockville, Maryland, USA: GSS, GHS, and ICF International. Available at: https://dhsprogram.com/pubs/pdf/FR307/FR307.pdf
Gonzalez, M.E., Ogus, J.L., Shapiro, G., and Tepping, B.J. (1975). Standards for discussion and presentation of errors in survey and census data. *Journal of the American Statistical Association.* 70 (351) 5-23.

Groves, R.M., Fowler Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2004). Survey Methodology. John Wiley & Sons., Hoboken, NJ.

Guillen, W.A. (2021). (Philippines) Adapting household survey data collection in challenging situations: The Philippine experience. Joint ISWGHS-UNESCAP Webinar: Covid-19 Assessment.

Harter R., Battaglia M. P., Buskirk T. D., Dillman D. A., English N., Fahimi M., Frankel M. R., Kennel T., McMichael J. P., McPhee C. B., Montaquila J., Yancey T., Zukerberg A. L. (2016), "AAPOR Report: Address-Based Sampling," Available at: https://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx#SECTION%203.

Hartleib, S., Langer, V., and Moser, C. (2021). Implementing CAWI in the Austrian Microcensus/LFS. LAMAS Workshop on Multi-Mode Data Collection.

Iacus, S. M., King, G. and Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. Political Analysis, Winter 2012, Vol. 20, No. 1, pp. 1-24.

INEC - Panamá (2020a). Encuesta de Mercado Laboral Telefónica. https://www.inec.gob.pa/publicaciones/Default3.aspx?ID_PUBLICACION=1037&ID_CATEGORIA=5&ID_SUBCATEGORIA=38

INEC - Panamá (2020b). September 2020 Telefónica Labour Market Survey Estimates Reliability.
https://www.inec.gob.pa/publicaciones/Default3.aspx?ID_PUBLICACION=1060&ID_CATEGORIA=5&ID_SUBCATEGORIA=38

Inter-Secretariat Working Group on Household Surveys (ISWGHS) (2020). Planning and Implementing Household Surveys Under COVID-19. Available at: https://unstats.un.org/iswghs/news/docs/COVID-19_TechnicalGNote_final.pdf

Inter-Secretariat Working Group on Household Surveys (ISWGHS, 2022a). Dashboard of COVID-19 Impact Surveys. Available at https://unstats.un.org/iswghs/task-forces/covid-19-and-household-surveys/COVID-19-impact-surveys/

Inter-Secretariat Working Group on Household Surveys (ISWGHS, 2022b). Positioning Household Surveys for the Next Decade. Available at https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji220042

Jablonski, W. (2014) Questionnaire design in telephone surveys: interviewers' and research call center managers' experience. University of Lodz
Jäckle, A., Lynn, P., Campanelli, P., Nicolaas, G., Hope, S. and Nandi, A. (2011). Causes of mode effects on survey measurement. Presented at the Fourth Conference of the European Survey Research Association (ESRA). Lausanne, Switzerland.

Jong, J. (2016). Telephone Surveys. Guidelines for Best Practice in Cross-Cultural Surveys. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved January 23, 2022, from http://ccsg.isr.umich.edu/

Kim, S-Y., Lim, K. E., and Lee, T.J. (2021). (Korea) Using mixed modes for household surveys amid COVID-19 – Lessons and implications from South Korea.

King, K., Petroni, R., and Singh, R. (1987). Quality Profile for the Survey of Income and Program Participation. U.S. Census Bureau Working Paper number SEHSD-WP1987-07 or SIPP-WP-30.

Kreuter, F., 2013. Improving Surveys with Paradata: Analytic Uses of Process Information. John Wiley & Sons

L'Engle K, Sefa E, Adimazoya EA, Yartey E, Lenzi R, et al. (2018) Survey research with a random digit dial national mobile phone sample in Ghana: Methods and sample quality. PLOS ONE 13(1): e0190902. https://doi.org/10.1371/journal.pone.0190902

Lessler, J., Tourangeau, R., and Salter, W. (1989). Questionnaire design in the cognitive research laboratory. Vital and Health Statistics 6:1, Hyattsville, MD. file:///Users/davidmarker/Downloads/cdc_11198_DS1.pdf

Lohr, S.L. (2022). Sample design and analysis, 3rd edition. Routledge Taylor & Francis Group. ISBN 9780367279509

Lynn, P. (2009). Methodology of Longitudinal Surveys. Wiley.

Mangiafico, S.S. (2016). Summary and Analysis of Extension Program Evaluation in R, version 1.19.10. rcompanion.org/handbook/

Marker, D.A. (2017). How have National Statistical Institutes improved quality in the last 25 years? *Statistical Journal of the International Association of Official Statistics* 33, 951-961.

Morgan, D. (2022). Making everyone count: how we're transforming the Labour Force Survey. Blog available at https://blog.ons.gov.uk/2022/03/29/making-everybody-count-how-were-transforming-the-labour-force-survey/

Morganstein, D. and Marker, D. (1997): Continuous Quality Improvement in Statistical Organization, In: Survey Measurement and process Quality, Lyberg, L.; Biemer, P.; Collins, M.; de Leeuw, E.; Dippo, C.; Schwartz, N. and D. Trewin (eds), New York: Wiley, pp. 475 – 500.

Nicolaas, G., Campanelli, P., Hope, S., Jäckle, A., and Lynn, P. (2011). Is it a good idea to optimize question format for mode of data collection? Results from a mixed modes experiment. Institute for Social and Economic Research.

Norwood, J.L. and Tanur, J.M. (1994). Measuring Unemployment in the Nineties. *Public Opinion Quarterly*. Vol. 58, No. 2, pp. 277-294. https://doi.org/10.1086/269424
Ochieng, Z.O. (2021). Discussion with Kenyan statistical office.

Olson, K., Smyth, J.D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N., McCarthy, J.S., O'Brien, E.O., Opsomer, J.D., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z.T., Turakhia, C., and Wagner, J. (2021). Transitions from Telephone Surveys to Self-Administered and Mixed-Mode Surveys: AAPOR Task Force Report. *Journal of Survey Statistics and Methodology*. P.397

Pew Research Center, January, (2018). "For Weighting Online Opt-In Samples, What Matters Most?"

Reese, D., Scanniello, N., and Ross, C.V. (2021). Adapting the American Community Survey amid COVID-19. Census Bureau Random Samplings Blog. https://www.census.gov/newsroom/blogs/random-samplings/2021/05/adapting-the-acs-amid-covid-19.html

Schork, J., Riillo, C.A.F., and Neumayr, J. (2021). Mixed Mode Effects of Web and Telephone Surveys Using Coarsened Exact Matching to Explore the Results on Employment Status. LAMAS Workshop on multi-mode data collection

Schouten, B., van den Brakel, J., Buelens, B., Giesen, D., Luiten, A., and Meertens, V. (2021). Mixed-Mode Official Surveys.

Shimizu, M. (2021). Development of training on phone survey for Asia and the Pacific. Joint ISWGHS-World Bank ESCAP75 Webinar.

Smith, Tom W. (2009). A revised review of methods to estimate the status of cases with unknown eligibility. American Association for Public Opinion Research. Available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/ERATE09.pdf

Smyth, J.D., Dillman, D.A., Christian, L.M. and Stern, M. J. (2006a). Comparing check-all and forced-choice question formats in web surveys. Public Opinion Quarterly, 70, 66-77.

Smyth, J.D., Dillman, D.A., Christian, L.M. and Stern, M. J. (2006b). Open ended questions in telephone and web Surveys. Presented at the World Association of Public Opinion Research Conference, May 16-18, Montreal, Canada.

Statistics Canada. (2021) Effects of the COVID-19 pandemic on data analysis and comparability over time: consideration for Canada and Census 2021. Presentation made during United Nations Second United Nations Expert Group Meeting on the Impact of the Covid-19 Pandemic on Conducting Population and Housing Censuses and on Census Data Quality Concerns, 2-5 November 2021. Available at https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Funstats.un.org%2Funsd%2Fdemographic-social%2Fmeetings%2F2021%2Fegm-covid19-census-20211102%2Fdocs%2Fs06-01-CAN.pptx&wdOrigin=BROWSELINK

Statistics Canada. (2022a). Analytical Guide - Portrait of Canadian Society: Perceptions of Life during the Pandemic. Available at https://www.statcan.gc.ca/en/statistical-programs/document/5352_D1_V1

Statistics Canada. (2022b). Summary of the Evaluation of Statistics Canada's COVID-19 Data Response: Crowdsourcing and Probability Panel Products and Specific COVID-19 Related Health Statistics. Available at: https://www.statcan.gc.ca/en/about/er/2021coviddr-oct-summary

Taskinen, P. (2021). LAMAS workshop on multi-mode data collection National experience: Finland.

Tavan, C. (2021). LAMAS Workshop Multi-mode data collection: The French experience

Torsteinsen, A. (2021). Statistics Norway's mixed-mode LFS pilot.

Tortora, R. (2004). Response trends in a national random digit dial survey, Advances in Methodology and Statistics, 1, 21-32.

Tourangeau, R. and Smith, T.W. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly*. 60:2. Pp. 275-304. https://www.jstor.org/stable/2749691

UNICEF (2021). Mongolia MICS Plus survey: Survey methodology.

United Nations (1964). Recommendations for the preparation of sample survey reports (provisional issue). Statistical papers, Series C No. 1, Rev. 2. New York: United Nations. https://unstats.un.org/unsd/publication/SeriesC/SeriesC_1_rev2.pdf

United Nations (2005). Household Sample Surveys in Developing and Transition Countries. Available at https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Handbooks/surveys/seriesf_96-E.pdf

United Nations (2019). United Nations National Quality Assurance Frameworks Manual for Official Statistics. Available at https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual/

United Nations Statistics Division (2020). Using telephone interviews for household surveys: A conversation with Prof. Jim Lepkowski. Available at: https://covid-19-response.unstatshub.org/posts/using-telephone-interviews-for-household-surveys/

United Nations Statistics Division and World Bank (2020a). Monitoring the State of Statistical Operations under the COVID-19 Pandemic, highlights from the third round of a global COVID-19 survey of National Statistical Offices (NSOs). Available at: https://covid-19-response.unstatshub.org/survey/covid-19-nso-survey-report-3.pdf

United Nations Statistics Division and World Bank. (2020b). Monitoring the State of Statistical Operations under the COVID-19 Pandemic, highlights from the first round of a global COVID-19 survey of National Statistical Offices (NSOs). Available at https://covid-19-response.unstatshub.org/survey/covid-19-nso-survey-report-1.pdf

United Nations (2021). Global SDG Data Platform. https://unstats.un.org/sdgs/unsdg. Data extracted in December 2021.

United Nations Economic Commission for Europe (UNECE) (2019). Generic Statistical Business Process Model, GSBPM Version 5.1. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

United Nations Economic Commission for Latin America and the Caribbean (UNECLAC) (2020a). COVID-19 Reports: Continuity of household surveys after the coronavirus disease (COVID-19) pandemic.
https://repositorio.cepal.org/bitstream/handle/11362/46521/1/S2000848_en.pdf

United Nations Economic Commission for Latin America and the Caribbean (UNECLAC) (2020b). Recommendations for eliminating selection bias in household surveys during the coronavirus disease (COVID-19) pandemic.
https://repositorio.cepal.org/bitstream/handle/11362/45553/1/S2000315_en.pdf

United Nations Economic Commission for Latin America and the Caribbean (UNECLAC) (2020c). Recommendations for the publication of official statistics based on household surveys in the face of the conjuncture of the coronavirus disease (COVID-19). COVID-19 Reports. Santiago. Obtained from
https://rtc-cea.cepal.org/sites/default/files/rtc_connected/files/recomendaciones-encuestas-hogares-covid-final.pdf

United Nations Economic Commission for Latin America and the Caribbean (UNECLAC) (2022). Lessons and challenges of the COVID-19 pandemic for household surveys in Latin America. ECLAC Statistical Briefings. No. 6 July 2022. ISSN: 2788-5836.

United States Census Bureau, 2022. Household Pulse Survey Technical Documentation. https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html#phase1

Valentino, Nicholas A., Kirill Zhirkov, Sunshine Hillygus and Brian Guay (2021). "The Consequences of Personality Biases in Online Panels for Measuring Public Opinion." Public Opinion Quarterly 84(2), 446-468.

Valliant, Richard, Dever, Jill A., Kreuter Frauke (2018). Practical tools for designing and weighting survey samples. Springer, 2018.

Valliant, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples. Journal of Survey Statistics and Methodology, 8(2), 231–263. https://doi.org/10.1093/jssam/smz003

Van den Brakel, J.A., Smith, P.A., and Compton, S. (2008). Quality procedures for survey transitions – experiments, time series, and discontinuities. Survey Research Methods, Vol. 2:3, pp.123-141. doi: 10.18148/srm/2008.v2i3.68. https://ojs.ub.uni-konstanz.de/srm/article/view/68 .

Van den Brakel, Souren, M., and Krieg, S. (2021). Estimating monthly labour force figures during the COVID-19 pandemic in the Netherlands. Central Bureau of Statistics discussion paper.

World Bank (2020). Papua New Guinea High Frequency Phone Survey on COVID-19: Results from Round 1. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/34907 License: CC BY 3.0 IGO.

Zalkalne, S. (2021). CAWI in Latvian LFS. LAMAS Workshop on multi-mode data collection.