IPRPD
International Journal of Arts, Humanities and Social Sciences
ISSN 2693-2547 (Print), 2693-2555 (Online)
Volume 07; Issue no 07: July 2026
DOI: 10.56734/ijahss.v7n7a6

FROM SAMPLING TO CENSUS: RETHINKING SURVEY METHODOLOGY IN THE DIGITAL ERA

Renier Steyn[1], Sean McCallaghan[2]*, Takawira Munyaradzi Ndofirepi[3], Angelo Joseph[4], Olanrewaju Adediran[5]

[1 2 3 4 5] *Graduate School of Business Leadership, University of South Africa (UNISA)*

Abstract

This article critically examines the historical reliance on sampling in survey research and argues for a renewed consideration of census-based approaches in the digital era. Traditional benchmarks have guided generations of researchers in determining the minimum sample sizes required. While these conventions remain influential, the growing availability of electronic sampling frames and administrative datasets increasingly makes census studies methodologically superior, as well as practical and cost-effective. The rise of Big Data further blurs the distinction between sampling and census research, enabling analyses of entire digital populations while introducing new methodological and ethical challenges related to representativeness, data quality, and privacy. The paper examines the economic, practical, statistical, and ethical benefits of census research, including higher response rates, reduced sampling error, the ease of post-stratification, and increased inclusivity. Attention is also given to the persistent challenge of non-response bias, with evidence suggesting that response rates, whether for samples or censuses, remain broadly comparable. Recommendations emphasise the adoption of census strategies where feasible, supported by digital infrastructure. Future research should investigate response dynamics and consider the ethical benefits associated with census approaches. Collectively, this analysis calls for adopting census designs and rethinking survey methodology, taking into account contemporary technological advancements.

Keywords

Census research; Sampling theory; Survey methodology; Response rates; Representativeness; Digital data collection; Big Data; Research ethics

Introduction

For decades, survey methodology has been dominated by sampling conventions. Generations of researchers have been trained to calculate minimum sample sizes, defend margins of error, and uphold rules of thumb such as the "rule of 30" (Lakens, 2022) or the benchmark of 400 respondents (Girish & Lee, 2020). These practices, grounded in classical statistics, continue to structure much of contemporary social science research and persist despite a radically changed research environment. In the digital age, access to administrative datasets, electronic records, and online platforms increasingly enables researchers to reach entire populations directly. Under these conditions, sampling is not only less essential, but also less justifiable than a census-based approach. This article argues that the entrenched reliance on sampling is outdated and that census approaches should be prioritised wherever population access is feasible. By examining the methodological, economic, practical, and ethical advantages of census research, this paper calls for a reorientation of the survey methodology.

A Historic Perspective on Sampling

Young researchers in the social sciences, particularly those conducting quantitative survey research, often go to great lengths to determine the minimum sample size required to meet accepted statistical parameters. A standard guideline is that approximately 400 respondents are needed to achieve population estimates with 95% confidence and a ±5% margin of error. This rule of thumb is derived from the classical sample size formulas for proportions (Cochran, 1977) and has been widely adopted in the social sciences. Krejcie and Morgan (1970) further reinforced this convention by providing a widely cited table of required sample sizes, in which a sample of 384 is indicated as sufficient for large populations, typically rounded to 400 for practical application. Accordingly, a sample size of

400 is considered the minimum necessary for most large-scale survey research studies to achieve adequate precision.

At the same time, the Central Limit Theorem provides a more lenient benchmark for what may constitute an adequate sample size (Sidebotham & Barlow, 2024). The theorem suggests that when the sample size is at least 30, the sampling distribution of the mean tends to approximate normality, regardless of the population distribution (Field, 2018). This "rule of 30" is frequently cited in statistics education as the minimum threshold for applying many parametric tests with reasonable accuracy. While this benchmark does not guarantee the same level of precision in population estimates as the convention of 400 respondents, it highlights that, under certain conditions, smaller samples can still yield valid and interpretable results.

In addition to general guidelines such as 400 respondents or the Central Limit Theorem's "rule of 30," researchers working with psychometric instruments and multivariate techniques often rely on proportional rules of thumb. A commonly cited standard is to secure at least 15 responses per item or per construct in a scale. This heuristic is intended to ensure sufficient statistical power and stable parameter estimates in analyses such as exploratory and confirmatory factor analysis, regression models, and structural equation modelling. For example, a 20-item scale would ideally require a minimum of 300 participants to satisfy this rule. Hair et al. (2019) recommend such a ratio as a practical benchmark, especially in complex models, where larger samples are critical for achieving reliable results.

Importantly, the adequacy of a sample is not determined by size alone but also by the type of sampling employed. Probability-based methods, such as simple random or stratified sampling, provide more representative data and allow generalisation to the wider population, whereas non-probability methods, such as convenience or snowball sampling, may limit external validity regardless of the sample's numerical size (Fowler, 2014; Field, 2018). Before moving forward, it is also important to consider the population from which the sample is drawn.

In the practical realities of research, access to populations is often constrained, and decisions about the sampling frame – the population from which participants are drawn – are typically shaped by convenience, cost, and feasibility. For instance, sampling frames are more often defined in urban than in rural contexts, and it is generally more practicable to collect data within one's own organisation than within another. Moreover, research populations of limited scope are frequently used to make generalisations or to build theory beyond the specific group studied. Historically, this has meant relying on so-called microcosms – for example, university students as stand-ins for broader populations, or even for humanity in general (Sears, 1986; Henrich, Heine, & Norenzayan, 2010), and employees of a single large multinational corporation as a proxy for the broader workforce (Hofstede, 1980). This observation is not a critique of these seminal works, but rather a reminder to those evaluating research proposals that young scholars face the same pragmatic constraints as their predecessors. Consequently, the sampling frames available to researchers – such as psychology students at Stanford or employees at IBM – have often provided the starting point for influential programmes of research. As such, access to the entire population is very seldom achieved.

In such research, the population is often conveniently selected, after which random or stratified sampling techniques are applied, with stratification employed to enhance representativeness. In fact, stratification is used to ensure that key subgroups within the accessible population are proportionally reflected in the sample. Researchers often go to great lengths to describe and justify these procedures, frequently at the expense of significant time and effort. Probability sampling is usually praised for its ability to produce generalisable results, particularly when implemented with rigour and when the cost of surveying the entire population is prohibitive (Saunders et al., 2019; Bryman, 2021).

To conclude, sample adequacy reflects a balance between statistical benchmarks, sampling methods, and practical constraints. While a sample of around 400 respondents is conventionally recommended to achieve a 95% confidence level with a ±5% margin of error (Krejcie & Morgan, 1970; Cochran, 1977), smaller thresholds are also defensible, such as the Central Limit Theorem's guideline of 30 cases (Field, 2018) or proportional rules requiring 15 responses per item in multivariate analyses (Hair et al., 2019).

The type of sampling further shapes adequacy, with probability methods supporting representativeness and non-probability methods limiting generalisability despite feasibility advantages (Fowler, 2014). Given that access to populations is often restricted, many seminal studies have relied on convenience samples such as university students or multinational employees (Sears, 1986; Hofstede, 1980), underscoring the need for transparency and justification in research design.

In the contemporary research environment, identifying sampling frames from which data can be drawn is often not especially difficult. Government institutions, such as the Department of Home Affairs, maintain extensive databases of email addresses and (smart) phone numbers linked to social media accounts, which could, in principle, provide access to large segments of the population. Access to such databases is, however, tightly regulated – and rightly so – to prevent misuse and manipulation. In more manageable contexts, researchers may instead gain access to defined populations, such as psychology students at a university like Stanford or employees of large multinational corporations, such as IBM. These data sources are also subject to the authority of institutional gatekeepers and governed by legislative requirements.

When delineating the research frame, survey participation is often restricted to individuals with email accounts or access to (smart) phones. Typically, this implies a minimum level of literacy and language proficiency, a requirement commonly met by those with institutional e-mail addresses provided by their university or employer. Employees without such access are generally excluded, as they are unlikely to meet the basic literacy skills necessary to complete written surveys effectively (Bryman, 2021). This helps to justify the use of electronic means of collecting data.

Now, equipped with a sample frame that can be translated into the available population, the next step is to apply a sampling technique. With simple random sampling, it is straightforward to draw a sample from a list of email addresses, for example, by using the RAND function in Excel or an online tool such as Research Randomiser, both of which ensure that each individual has an equal probability of selection (Fowler, 2014).

If stratification is to be applied to enhance representativeness, it may be challenging when relying solely on e-mail addresses or phone numbers, as these identifiers typically do not include relevant demographic information, such as gender or age. Stratification is therefore only feasible when contact lists are linked to auxiliary data, such as institutional or organisational records, or when demographic information is collected through preliminary screening questions.

Sampling in the 21st Century

Sampling – and the many hours devoted to it – was central in the years before the emergence of the electronic age. However, when an electronic sampling frame is available, conducting a full census rather than a survey may offer advantages.

The first aspect to consider is sampling error. When information is collected from every member of the population, as in a census, sampling error is eliminated, and the resulting statistics represent the true population parameters. In such cases, inferential statistics and confidence intervals are unnecessary, as no generalisation from the sample to the population is required. Instead, the observed statistics directly describe the characteristics of the population and the relationships between variables (Cochran, 1977; Bryman, 2021).

The next challenge is non-response. Conceptually, the population consists of both respondents and non-respondents, and two aspects are essential to consider: the absolute size of the non-response group and the characteristics of those who did not participate. Generally, response rates for online surveys are significantly lower than those obtained in traditional paper-and-pencil surveys (Nulty, 2008; Sauermann & Roach, 2013). It may also be argued that organisational surveys achieve somewhat higher response rates than general online surveys. Yet, they carry a different limitation: participants in paper-based or organisational contexts may be more vulnerable to coercion or peer pressure, whereas online environments provide greater freedom to decline participation. This creates an interesting position: although online surveys typically achieve lower response rates, they also yield data that is provided freely, without coercion, and in a manner consistent with ethical standards. Nevertheless, non-response remains a serious concern. Research shows that non-respondents are not always randomly distributed across the population; they often differ systematically from respondents in demographic characteristics, attitudes, or behaviours (Groves & Peytcheva, 2008). Such differences can introduce non-response bias, thereby undermining the representativeness and validity of the findings. Moreover, when non-respondents substantially outnumber respondents, the greater concern becomes responder bias, as those who choose to participate may differ markedly from the broader population. In this sense, the issue lies not only in the absence of non-respondents but also in the selective nature of respondents, whose motivations or traits may not be evenly distributed across the population (Groves & Peytcheva, 2008; Bethlehem, 2010).

When all possible respondents are accessible through established channels (e.g., government websites or institutional e-mail), conducting a census offers a distinct administrative advantage. It leverages existing systems to reach the entire population, thereby reducing the need for additional resources devoted to sampling design and management (Hair et al., 2023). Moreover, with the rise of online data collection tools and platforms, the cost and logistical barriers to conducting complete censuses in organisations have been significantly reduced (Creswell & Creswell, 2018).

Considering a census, which includes all possible respondents rather than a sample, carries important ethical and methodological advantages. A census has the potential to enhance inclusivity and help avoid stratification biases that may arise in sampling approaches. Inclusivity is a cornerstone of research ethics, as it ensures that every individual has the opportunity to contribute their perspective to the debate, the formulation of solutions, and even the design of interventions or participation in experiments. In this way, no potential participant can reasonably ask, "Why was I not included?" since all members of the defined population are given an equal voice (Bryman, 2021; Saunders, Lewis, & Thornhill, 2019).

Latimore and Lewis (2017) examined this issue in the context of large-scale U.S. federal employee surveys. They found that organisations often abandon sampling when the sample encompasses a large proportion of the total workforce, opting instead for a census strategy. Their findings indicated that identifying respondents as part of a census did not negatively affect participation, suggesting that census approaches are not inferior in

practice for attracting respondents. Put differently, whether individuals know they were specifically selected or realise that they are part of the entire population appears to have little effect on response rates. Response rates for online surveys, whether based on a sample or a census, appear to be broadly comparable.

Stratification within a census can also be applied with relative ease. In the case of a sample, stratification must be designed prior to data collection but in a census, stratification can be implemented after the fact by using available demographic variables to create categories. If necessary, strata can then be weighted to compensate for low response rates in certain groups (Bethlehem, 2010). For example, if it is known that 50% of the workforce is based in the United States but only 20% of respondents are from that group, such discrepancies may not only be adjusted statistically but can also provide insight into patterns of non-response.

Perhaps most important of all is that a census will almost always yield more respondents than a sample survey. For example, if you send out 400 surveys as part of a sampling process to a population of 3,000, you may, if fortunate, end up with around 300 completed responses. By contrast, if you send surveys to all 3,000 individuals in the population, you are likely, given the same response rate, to receive far more than 400 responses. This provides two key advantages: first, it enables the use of statistical techniques that require larger sample sizes; and second, the larger pool of respondents allows for stratification and weighting, thereby supporting a more balanced representation of the population.

The Role of Big Data

The emergence of Big Data has increasingly blurred the distinction between sampling and census research. Big Data encompasses vast, continuously generated digital datasets derived from online transactions, social media interactions, sensor networks, and administrative systems. Characterised by their volume, velocity, and variety, such data enable analyses that approximate or even exceed the scope of traditional censuses (De Mauro et al., 2016). Rather than relying on samples, researchers can now access entire digital populations, such as all users of a specific platform or all entries within an administrative registry, thereby achieving a de facto census (Crawford et al., 2014). This evolution redefines what constitutes a population in empirical research, as computational tools and analytical capacity increasingly permit the examination of entire systems rather than subsets thereof (Kitchin & Lauriault, 2018).

Nevertheless, Big Data presents substantial methodological and ethical challenges, particularly regarding representativeness, data quality, and algorithmic bias. Accordingly, Big Data research must be firmly grounded in contextual awareness, encompassing data sources, regulatory frameworks, and potential conflicts of interest. While findings derived from such data may appear conclusive, the origins of the data are often complex and fragmented, necessitating careful attention to provenance and limitations to ensure accurate and responsible interpretation (boyd[1] & Crawford, 2012; Zook et al., 2017). Furthermore, questions of privacy, consent, and fairness remain central to the ethical governance of large-scale datasets (Taylor & Broeders, 2015). Integrating Big Data into survey methodology, therefore, demands rigorous ethical scrutiny and methodological reflexivity. Kitchin (2014) warns of consequences beyond immediate results, and states that Big Data contributes to the "creation of data-driven rather than knowledge-driven science, and the development of digital humanities and computational social sciences that propose radically different ways to make sense of culture, history, economy and society."

Recommendations

If the aim is to benchmark a conventional standard for survey precision at a 95% confidence level and ±5% margin of error, the rule of thumb is to use a sample size of 400. Ultimately, all calculations culminate in this point, eliminating the need for an explicit formula. The exception may apply to multivariate analyses such as factor analyses, where proportional rules requiring 15 responses per item apply (Hair et al., 2019), potentially necessitating larger sample sizes. The main point is that relatively large samples are required for generalisation, whereas a census addresses this issue directly. In a census, there is no need for generalisation, since the entire population is surveyed. As a result, confidence intervals and margins of error become unnecessary, because the data represent true population parameters rather than estimates. Thus, it is recommended that, where possible, samples be abandoned and census data collected instead.

Use a census rather than sampling for economic reasons. There is little more required than obtaining the respondents' particulars and then proceeding with the research. Where full access to the population is available, few administrative issues arise.

[1] Stylised in lowercase, danah boyd.

Use a census rather than sampling for practical reasons. It is relatively easy to access datasets containing contact details, and if stratification is required, it can be achieved by including stratifying questions within the survey. A census will always yield more responses than a sample.

Use a census for statistical reasons. Conducting a census eliminates sampling error, meaning no inferences are required. In addition, stratification and weighting can be applied after data collection – a process known as post-stratification – which is relatively straightforward.

Use a census for ethical reasons. Census approaches aim to avoid exclusion and ensure that all members of the population have an opportunity to participate, thereby upholding the principles of inclusivity in research ethics.

Use a census with the rule of thumb that if a 50% response rate is achieved, the census may be considered reasonably representative of the population. Although this benchmark is arbitrary, it addresses concerns that non-respondents may differ systematically from respondents. Once response rates approach 50%, the non-respondent group becomes so large, relative to the respondents, that the resulting data may reasonably be viewed as indicative of the population.

Use Big Data only where its context, provenance, and limitations are clearly understood and reported. While Big Data offers the potential for near-census coverage and rich analytical insights, it should be employed only when researchers can ensure the quality, representativeness, and ethical integrity of the datasets involved. Its use demands careful consideration of data sources, collection methods, and possible biases, as well as transparency regarding these factors.

In conclusion, and in addressing the earlier question of the extent to which census-based methods should replace sampling in the context of digital-era research, the answer is unequivocal: census approaches should prevail. There is little, if any, justification for not adopting such a practice wherever feasible.

Conclusion

Although established benchmarks and practices in sampling continue to guide social science research, future studies should examine how these conventions adapt to evolving research contexts. First, more empirical work is needed to compare the effectiveness of census and sampling approaches in different organisational and societal settings, particularly as digital technologies make full-population surveys more feasible. What needs to be considered is who will be excluded if such online studies are conducted. Second, future research should investigate strategies to mitigate non-response bias in online surveys, including the role of incentives, survey design, and follow-up methods. Linked to this is the proposed rule of a 50% response rate. How well does a response rate of 50% represent the entire population?

Third, methodological studies should assess the extent to which proportional rules of thumb (e.g., 15 responses per item) hold across increasingly complex analytical techniques, such as multi-level modelling used when testing for measurement invariance. Large groups are needed for such research. Finally, the ethical implications of inclusivity in sampling, as well as in "imperfect" censuses and Big Data, require greater attention. The processes involved in data creation, access to administrative records, privacy protection, and the securing of informed consent demand careful and transparent consideration. These ethical challenges persist and remain unresolved, regardless of how the data are produced, accessed, or applied. By pursuing these avenues, future research can ensure that sampling theory and practice remain robust, relevant, and responsive to the demands of contemporary social science.

Declarations

Conflict of interest: No potential competing interest was reported by the authors.
Data availability statement: Data sharing not applicable – no new data generated.

Generative Artificial Intelligence (AI): ChatGPT (December 2024 version) was used as an aid to improve sentence construction and grammar in the text.

Authors' contributions: The first author was responsible for the conceptualisation and structural design of the manuscript and prepared a rough draft. All authors contributed by reviewing and editing the draft, providing critical feedback that enhanced the clarity and depth of the work. All authors read and approved the final version of the manuscript prior to submission.

Data availability statement: No new data were generated or analysed for this study; all information was sourced from previously published material.

Ethical considerations: This is a conceptual paper.

Funding: This research did not receive any sponsorship or funding.

Acknowledgement: Lydia von Wielligh-Steyn edited the manuscript.

References

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review, 78*(2), 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x

boyd, d., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bryman, A. (2021). *Social research methods* (6th ed.). Oxford University Press.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing Big Data: Politics, ethics, epistemology. *International Journal of Communication, 8*, 1663–1672.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). What is big data? A consensual definition and a review of key research topics. *Journal of Big Data, 2(*1), 1-21.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.

Fowler, F. J. (2014). *Survey research methods* (5th ed.). Sage.

Girish, V. G., & Lee, C. K. (2020). Authenticity and its relationship with theory of planned behaviour: Case of Camino de Santiago walk in Spain. *Current Issues in Tourism, 23*(13), 1593-1597. https://doi.org/10.1080/13683500.2019.1676207

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly, 72*(2), 167–189. https://doi.org/10.1093/poq/nfn011

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.

Hair, J. F., Page, M., & Brunsveld, N. (2023). *Essentials of business research methods* (6th ed.). Routledge.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Sage.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1–12. https://doi.org/10.1177/2053951714528481

Kitchin, R., & Lauriault, T. P. (2018). Towards critical data studies: Charting and unpacking data assemblages and their work. In *Thinking big data in geography: New regimes, new research*, pp. 3-20. University of Nebraska Press.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*(3), 607–610. https://doi.org/10.1177/001316447003000308

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology, 70*(2), 144.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1), 33267. https://doi.org/10.1525/collabra.33267

Latimore, L., & Lewis, T. H. (2017). Are response rates to organizational climate surveys influenced by informing respondents about the sampling strategy? *Survey Practice, 10*(2). https://doi.org/10.29115/SP-2017-0004

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education, 33*(3), 301–314. https://doi.org/10.1080/02602930701293231

Sauermann, H., & Roach, M. (2013). Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features. *Research Policy, 42*(1), 273–286. https://doi.org/10.1016/j.respol.2012.05.003

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research methods for business students* (8th ed.). Pearson Education.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and* Social Psychology, 51(3), 515–530. https://doi.org/10.1037/0022-3514.51.3.515

Sidebotham, D., & Barlow, C. J. (2024). The central limit theorem: The remarkable theory that explains all of statistics. *Anaesthesia, 79*(10). https://doi.org/10.1111/anae.16420

Zook, M., Barocas, S., boyd, d., Crawford, K., Keller, E., & Gangadharan, S. P. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology, 13*(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399