The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2531-0488.htm

ThinkBox

Using the internet for data collection in business research

Mateus Canniatti Ponchio
Escola Superior de Propaganda e Marketing – ESPM, Sao Paulo, Brazil, and
Nelson Lerner Barth and Felipe Zambaldi
School of Business Administration in Sao Paulo, Fundacao Getulio Vargas, Sao Paulo, Brazil

Introduction

Imagine a master's degree student facing the need to collect primary data to proceed into quantitative analysis in the business field. The tight deadline for getting the degree and the inherent difficulty in research planning and collecting data can both be frightening. Before the internet, the alternative would be a convenience non-probabilistic sampling, which is a generally easier and faster way to obtain data than probabilistic sampling. Thus, considering the internet's use to speed up data collection and reduce costs, what are the advantages and recommendations to be accounted for by business researchers?

The use of the internet to conduct surveys is attractive for business researchers due to its benefits in speed, access to respondents and the large amounts of data it can gather. The main advantages of online versus offline data collection are time-saving, cost reduction, simplified data tabulation and purification processes, flexibility and format control. However, there are issues related to respondents' attention, sample representativeness and control, which can be amplified as most of the data collection for business research on the internet occurs through convenience sampling. Also, self-selection bias can represent severe jeopardy in online surveys.

In this short essay, we draw on the problem of sample representativeness on Web-based data collection, and then we discuss it with emphasis on two common procedures to mitigate its risks and challenges:

- the correction of snowball sampling bias; and
- the use of respondent panels.

The reflection is structured in four additional sections. Firstly, we introduce the issues related to convenience sampling size and the problem of representativeness. We then

© Mateus Canniatti Ponchio, Nelson Lerner Barth, Felipe Zambaldi, Mateus Canniatti Ponchio, Nelson Lerner Barth and Felipe Zambaldi. Published in RAUSP Management Journal. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence maybe seen at http://creativecommons.org/licences/by/4.0/legalcode.

RAUSP Management Journal
Vol. 56 No. 2, 2021
pp. 254-258
Emerald Publishing Limited
2531-0488
DOI 10.1108/RAUSP-04-2021-251

discuss approaches to correct sampling bias and the use of respondent panels from professional market research suppliers. Finally, we bring further caution-related aspects of conducting online data surveys and present other data collection issues for business research on the internet and their need for future reflections.

Large convenience sampling and representativeness

There is a common misleading belief that if a sample is large (perhaps with millions of responses), it is automatically proper for inference about a population of interest, which could avoid the need to obtain a probabilistic sample. One can indeed remove the elements that do not belong to the sample's target population (e.g. non-buyers, if the target population is buyers only). However, even if the sample size has practically the same order of magnitude of the population, sampling biases may remain as there can be certain profiles that are over-represented or under-represented.

One can try to correct the proportions observed in the sample by weighting the actual populational proportion, which is called quota sampling (Rukmana, 2014). For example, if we admit (fictitious number) that 52% of a population comprises females, we can balance the sample to respect this percentage. We can apply the same solution to numerous other easily observable variables (age and type of housing). However, it is delusional to think that this method can eliminate the bias of over- or under-representation of specific profiles, as there are countless other variables for which we cannot adjust the sample's proportions (consumer behaviours, convictions, sensitivities, aversions, attractions and wishes). In other words, quota sampling is not an adequate substitute for probabilistic sampling when we need statistical inferences for a targeted population.

Out of the convenience sampling techniques that allow for obtaining large samples on the internet, snowball is quite popular due to its easiness to reach respondents. In the snowball sampling technique (Heckathorn, 2011), a group of people is invited to respond to questionnaires, usually using social networks and/or emails. The researchers ask each respondent to pass on the invitation to other people while offering rewards to the participants (that can be a gift, cash value or merely the prestige and pleasure of belonging to the group). But even if it is possible to restrict the sample to the target audience of the research, this technique can generate strong bias in its results. Two of these possible effects are there is no randomisation in the choice of the initial respondents, and there is no randomisation when respondents invite other respondents from their networks. There are, though, advantages in the use of snowball sampling and possible corrections for its bias, which we discuss next.

Snowball sampling bias and corrections

Digital social networks facilitate surveys to obtain data, as we can find and recruit large samples of respondents to apply questionnaires. Nevertheless, the use of the internet for sampling can be problematic because it may reproduce difficulties that were already there even before the internet. For example, textbooks have alerted us about convenience sampling problems, pitfalls of snowball techniques and other non-probabilistic sampling approaches (Churchill & Iacobucci, 2005; Pedhazur & Schmelkin, 1991). These are precisely the most common data collection approaches through the internet that we see in business research.

However, we should note that snowball sampling makes data collection more manageable, allowing access to typically censored groups (e.g. drug users). There are no previously available lists to select respondents from randomly. This fact encouraged the creation and development of statistical techniques to correct the inherent bias of sampling

methods. Such techniques belong to the category respondent-driven sampling (RDS) (Beaudry & Gile, 2020; Heckathorn, 2011). The non-probabilistic choice of respondents' bias is corrected based on the information passed on by the respondents themselves about the characteristics and size of their networks of contacts, admitting that there is a large number of successive invitations to new respondents.

Tools for applying RDS include RDSAT, based on Java (Volz et al., 2012), and R's RDS library (Handcock, Fellows, & Gile, 2019). Given the need for mastering specific statistical knowledge to use these techniques for snowball sampling bias correction – usually not the primary specialities of business researchers – the technique requires knowledgeable researchers in terms of these bias correction procedures or specialised professionals' support to run the corrections.

In addition to correcting sampling bias, one can avoid them by using sampling controls. There are professional market research suppliers that are skilled and offer respondent panel services. Next, we present the characteristics of these panels and some reflection on them.

Respondent panels

Many research firms provide respondent panels (a search on Google with the words "online research panel" returns dozens of companies). Usually, these companies recruit potential respondents to get registered in their databases, who get rewards to complete questionnaires. The cost for researchers to have access to these online panels can vary according to the target population (e.g. a survey with directors of multinational companies will be more expensive, per questionnaire, than a survey with bottled water consumers, as the former are harder to find and recruit). The probably most well-known platform for accessing respondents is Amazon Mechanical Turk (MTurk), a crowdsourcing service. Its users, such as organisations and researchers, can hire crowdworkers, remote outsourced respondents who perform the tasks of completing questionnaires.

A variety of studies have accessed the benefits of using MTurk (Barends & Vries, 2019; Cheung, Burns, Sinclair, & Sliter, 2017; Follmer, Sperling, & Suen, 2017). For instance, data quality assurance, as MTurk often includes attention-checking questions; nevertheless, researchers should include their own set of attention checks as well. A typical criticism regarding respondent panels is that, as crowdworkers, participants may tend to become professional respondents along time, thus losing their spontaneity in providing answers, which cannot be monitored by attention checks. On the contrary, they may become overfocussed on not missing the checks and guaranteeing their rewards or even trying to provide answers that they believe the researchers would like them to give. This behaviour might become a serious issue, as the most common situations in primary business research are those in which questionnaires require spontaneity from respondents.

For the present discussion, it suffices to say that, regardless of the media used to gather respondents, there must be a focus on the respondent panel's suitability to the research objectives. For example, not everyone has access to email or the internet. The global penetration of the internet is estimated to be roughly 60% at the end of 2020 (Statista Portal, 2020), which varies significantly according to regional infrastructure and development variability. Therefore, research objectives should determine the target population and, thus, the appropriateness of online data collection approaches.

Further recommendations for conducting online data collection

Web-based online data collection methods create opportunities to conduct research globally, especially among difficulties to access populations. It should be noted that, although online data collection brings cost- and time-related benefits, especially for geographically dispersed

populations, the internet partly reproduces and amplifies limitations that existed priorly. Web-based research requires careful consideration of how studies will be promoted and how data will be collected to ensure the validity of findings (Cantrell & Lupinacci, 2007; Lefever, Dal, & Matthíasdóttir, 2007; Topp & Pawloski, 2002). Caution should be the rule for defining a target population and the criteria to allow the drawing of an unbiased sample.

Online data collection can be carried out using different interfaces that require internet browsers or a smartphone application. Each approach has advantages and disadvantages, and the appropriate choice can save time when organising and analysing data. It is good practice to look for evidence of response rate (refusal rate) and selection bias. As people are overwhelmed with online information, an invitation to participate in a study may be skipped or consciously ignored, bringing distinct consequences to the research's response rate.

In this short essay, we have discussed concerns about using the internet and digital social networks to conduct surveys. However, other limitations persist and require new reflections, such as further issues regarding sample representativeness, computation of response rates and technical difficulties. Besides that, future business research needs to consider how to match the online environment's characteristics with common textbook recommendations for surveys. For example, defining study populations clearly, how to contact respondents, designing instruments for data collection while assuring face and content validity, pre-testing and adjusting, respondents' privacy and ethical aspects (Churchill & Iacobucci, 2005).

Finally, non-survey Web-based techniques for data collection, such as observing social networks, monitoring internet browsing behaviours and applications for gathering data in mobile devices, can provide researchers with relevant information but require further understanding. Ethnographic studies on digital and social networks, known as digital ethnography and cyber ethnography (Atay, 2020; Lester, 2020) or netnography (Kozinets, 2019) may imply obtaining data from digital social networks or the Web environment. This practice, in general, leads to conducting behaviour analyses without asking any questions. When the object of study is the population of individuals who use the digital social network, there are no issues related to poor sampling or lack of representativeness. There are, however, privacy and ethical concerns to be observed. As non-survey techniques to collect data on the internet tend to become more frequent in business research, the knowledge on them needs to be improved and further discussed by the business research community.

There is no one-size-fits-all answer to whether online data collection will be better or even suitable for a research study. The answer needs to consider the research questions and study objectives. The data source's quality cannot be defined by the data source itself, but rather by the researchers' knowledge and consequent decisions during the stages of study design, data collection and analysis.

References

Atay, A. (2020). What is cyber or digital autoethnography? *International Review of Qualitative Research*, *13*(3), 267–279. doi: https://doi.org/10.1177/1940844720934373.

Barends, A. J., & Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences*, *143*(1), 84–89. doi: https://doi.org/10.1016/j.paid.2019.02.015.

Beaudry, I. S., & Gile, K. J. (2020). Correcting for differential recruitment in respondent-driven sampling data using ego-network information. *Electronic Journal of Statistics*, *14*(2), 2678–2713. doi: https://doi.org/10.1214/20-EJS1718.

Cantrell, M. A., & Lupinacci, P. (2007). Methodological issues in online data collection. *Journal of Advanced Nursing*, *60*(5), 544–549. doi: https://doi.org/10.1111/j.1365-2648.2007.04448.x.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon mechanical Turk in organizational psychology: an evaluation and practical recommendations. *Journal of Business and Psychology*, *32*(4), 347–361. doi: https://doi.org/10.1007/s10869-016-9458-5.

Churchill, G. A., & Iacobucci, D. (2005). *Marketing research: methodological foundations*, 9th ed., p. 697. Mason, OH: Thomson.

Follmer, D. J., Sperling, R. A., & Suen, H. K. (2017). The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher*, *46*(6), 329–334. doi: https://doi.org/10.3102/0013189X17725519.

Handcock, M. S. Fellows, I. E., & Gile, K. J. (2019). RDS: Respondent-Driven sampling. R package version 0.9-2. Retrieved from: https://CRAN.R-project.org/package=RDS

Heckathorn, D. D. (2011). Snowball versus respondent-driven sampling. *Sociological Methodology*, *41*(1), 355–366. doi: https://doi.org/10.1111/j.1467-9531.2011.01244.x.

Kozinets, R. V. (2019). *Netnography: the essential guide to qualitative social media research*, SAGE Publications Limited.

Lefever, S., Dal, M., & Matthíasdóttir, A. (2007). Online data collection in academic research: Advantages and limitations. *British Journal of Educational Technology*, *38*(4), 574–582. doi: https://doi.org/10.1111/j.1467-8535.2006.00638.x.

Lester, J. N. (2020). Going digital in ethnography: Navigating the ethical tensions and productive possibilities. *Cultural Studies ↔ Critical Methodologies*, *20*(5), 414–424. doi: https://doi.org/10.1177/1532708620936995.

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design, and analysis: an integrated approach*, pp. 819. Hillsdale, NJ: Lawrence Erlbaum Associates. p

Rukmana, D. (2014). Quota sampling. In A. C. Michalos, (Ed.), *Encyclopedia of quality of life and Well-Being research*, Dordrecht: Springer.

Statista Portal. (2020). Global digital population as of October 2020. Retrieved from www.statista.com/statistics/617136/digital-population-worldwid

Topp, N. W., & Pawloski, B. (2002). Online data collection. *Journal of Science Education and Technology*, *11*(2), 173–178. doi: https://doi.org/10.1023/A:1014669514367.

Volz, E., Wejnert, C., Cameron, C., Spiller, M., Barash, V., Degani, I., & Heckathorn, D. D. (2012). *Respondent-driven sampling analysis tool (RDSAT) version 7.1*, Ithaca, New York, NY: Cornell University.

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**