BMJ Open

The pandemic of online research in times of COVID-19

Jeroen De Man ![ORCID],[1] Linda Campbell,[2] Hanani Tabana,[3] Edwin Wouters ![ORCID] [2,4]

¹Department of Family Medicine and Population Health, University of Antwerp, Antwerp, Belgium
²Centre for Population, Family and Health, Department of Sociology, University of Antwerp, Antwerp, Belgium
³School of Public Health, Faculty of Community and Health, University of the Western Cape, Cape town, South Africa
⁴Centre for Health Systems Research & Development, University of the Free State, Bloemfontein, South Africa

Correspondence to
Dr Jeroen De Man;
jeroen.deman@uantwerpen.be

ABSTRACT

The COVID-19 pandemic has led to an explosion of online research using rating scales. While this approach can be useful, two of the major challenges affecting the quality of this type of research include selection bias and the use of non-validated scales. Online research is prone to various forms of selection bias, including self-selection bias, non-response bias or only reaching specific subgroups. The use of rating scales requires contextually validated scales that meet psychometrical properties such as validity, reliability and—for cross-country comparisons—invariance across settings. We discuss options to prevent or tackle these challenges. Researchers, readers, editors and reviewers need to take a critical stance towards research using this type of methodology.

INTRODUCTION

Assessment of the biopsychosocial impact of the COVID-19 pandemic has been identified as a pressing need.[1] For a variety of reasons, including pragmatic ones, the COVID-19 pandemic has pushed many researchers to measure these biopsychosocial consequences using a particular methodological approach: the assessment of latent variables measured with rating scales in online questionnaires. This approach may indeed circumvent barriers in data collection if based on rigorous methods and a clear and appropriate research question. However, while using this approach initially seems straightforward, it has a long-standing track record of challenges and uncertainties, from setting up the research to the analyses and reporting of results.[2 3] The pressure to use online surveys in a short space of time may therefore be leading to another pandemic: one of studies with poorly constructed scales and non-representative sample sizes. In this commentary, we highlight possible drawbacks of the use of online surveys and rating scales, and we offer potential solutions for those who endeavour to use them.

Selection bias in online research

Sample validity is an essential requirement in survey research and means that each participant of the study population has the same chance to participate. Potential threats to sample validity are well known when using online studies.[2] First, where and how surveys will be made available will strongly determine their participants. Studies using distribution channels that only reach a subgroup of the target population suffer from selection bias.[4] Differences in health literacy and online access may strongly skew participation, especially in low-income and middle-income countries or in societies with large differences in educational and socioeconomic levels. Second, online surveys will typically attract participants who have a special interest or a close relationship with the topic (ie, self-selection bias).[4] On the other hand, specific subgroups may be less inclined to respond or complete the survey (ie, non-response bias).[4] When only a subgroup is being reached, selection bias will typically grow with an increasing diversity of the target population. This implies that extra caution is needed when studying more general populations.

Even though online surveys may obtain large sample sizes, this does not necessarily compensate for selection bias and may even make it worse.[5] Correction of such bias is often a daunting task, if possible at all. Suggestions to prevent it include the following:

1. Balance information in the introduction of the survey to sufficiently inform potential participants and to avoid eliciting interest from a particular subgroup. For example, mentioning that the survey relates to COVID-19 may be needed to attract sufficient participants, while introducing the survey as 'COVID-19 and your mental health status' may only attract a specific subgroup.
2. Include a broad array of items to measure sociodemographic and other characteristics that may potentially determine participation. Reporting these characteristics will help authors and readers to appraise sample validity and recognise the study's limitations. Using the same questions as in large surveys, such as a population census, may allow better assessment

of the sample's representativeness and may allow application of sample weighting. A recent study by De Man *et al* could serve as an example.[6] This study investigated associations between COVID-19-related stressors and depression in Belgian students attending higher education. When comparing their study sample with governmental data on higher education students enrolled in the previous year, they found a higher proportion of women (±20%), while other sociodemographic characteristics were comparable. This higher proportion of women needs to be taken into account when interpreting the findings of the study, especially since depression is more common among women.

3. How and to whom the survey is distributed is crucial. Potential distribution channels for online surveys include social media, news outlets, phone messages, email lists and quick response (QR) codes on printed material. Pursuing sample validity requires a tailored approach that facilitates equal participation of all relevant subgroups of the target population. For example, distributing the survey through academic networks will likely result in a very skewed image of the general national population. A recent study on the use of Facebook as a recruitment strategy tested an intervention to improve sample validity: the implementation of male-only advertisement increased the proportion of male participants.[7] However, obtaining a representative sample may sometimes not be possible, and researchers may need to rethink whether launching an online questionnaire is warranted. If researchers choose to go ahead, robust reporting of the study procedures in the methods and discussion section is essential. The Checklist for Reporting Results of Internet E-Surveys (CHERRIES) checklist may serve as a useful guide for this matter.[2]

4. Given the current mushrooming of online initiatives, respondents may feel overloaded by the sheer number of questionnaires they are presented with, thereby reducing their interest. Avoiding lengthy questionnaires can help. Joining forces with other research groups can reduce the number of duplicate initiatives and increase access to different distribution channels. In particular, for surveys on COVID-19, early registration of projects may facilitate such collaboration. A global COVID-19 research registry for public health and social sciences can be found here (https://converge. colorado.edu/resources/covid-19/public-health-social-sciences-registry).

5. Keeping participants informed about the results (eg, through the press and individual base) and presenting them as coproducers of knowledge may encourage participation in future initiatives. In the previously mentioned study by De Man *et al*, authors may consider giving feedback to their study participants through student associations, fraternities or the university communication.

The use of rating scales

If and how rating scales should be used in the assessment of latent variables has been subject of an ongoing debate.[8] A theoretical discussion on the type of data that rating scales represent (eg, continuous, ordinal and interval) is beyond the focus of this commentary. However, if one decides to use rating scales, current consensus converges on the need to meet specific psychometrical properties, especially if the results are used as unit-weighted composite scores (ie, the summation of scores per participant). In the following paragraph, we highlight some essential properties that are often neglected. For more details and psychometrical estimation methods, we refer to relevant textbooks or online resources such as the ones developed by Revelle.[9]

1. To draw meaningful conclusions, a scale needs to be valid for the studies' target population.
   Validity can refer to a scale's proven ability to predict a certain outcome (ie, predictive criterion validity) or a scale's relationship with a well-established measure or gold standard (ie, concurrent criterion validity). Contextual similarity between the actual study and the validation setting is essential and often overlooked.

2. Content validity corresponds to the scale measuring all facets of a given construct. A common and well-intended practice to reduce the length of the survey is to use a selection of the items of an existing scale. However, this shortening may also reduce construct coverage and consequently affect a scale's content validity.

3. For cross-validation purposes, researchers may also consider using scales that have been used in large representative surveys in their study setting (eg, the Demographic and Health Survey and the European Social Survey).

4. The use of unit-weighted composite scores is often justified by a coefficient alpha estimate above a defined cut-off depending on the purpose and field of research. First of all, it is important to note that a high value of alpha is not a sufficient criterion to use composite scores. Often overlooked is the participants' conceptual understanding of the items as has previously been noted. Theoretically, items that measure completely different concepts could correlate with each other and result in an acceptable alpha. In addition, testing reliability of a composite score is not straightforward as the validity of alpha depends on rather strict conditions.[10] One such condition is unidimensional data (ie, the scale measures only one concept). A scale that measures various subfactors besides an overarching 'general' factor may therefore overestimate coefficient alpha. In addition to alpha, the use of model-based estimators such as total omega and hierarchical omega is now recommended in order to arrive at a more nuanced estimation of the reliability of a scale.[10] In particular, if a scale shows deviations of unidimensionality (ie, the scale does not measure only one concept), reliability assessment becomes complex, and we would not do justice to formulate a simple guideline for its estimation. More details

on the recommended procedures to test reliability of a scale can be found in textbooks or online.[9] Finally, it is important to mention that, even though the use of composite scores may be justified based on reliability estimates, techniques such as factor analysis and structural equation modelling are usually preferred when it comes to accuracy of measuring relationships with one or more latent variables.

5. Comparison over time, across different settings (eg, different countries) and between subgroups, may be of special interest when studying latent constructs in participants exposed to rapidly changing environments because of COVID-19 or related preventive measures (eg, lockdown). While often overlooked, comparison of subgroups, be it through factor analysis or through composite scores, requires measurement equivalence, which can be defined as 'whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute'.[11] Impaired measurement equivalence precludes a meaningful interpretation of measurement data and can be due to various reasons.[12] For instance, different subgroups may attribute a different meaning to certain words of an item because of a different socioeconomic background or because of a different language use: the interpretation of 'feeling stressed' may substantially differ among countries or even between men and women.

6. Depending on the purpose of the study, measurement equivalence can be tested through different levels of measurement invariance. For instance, if the relationship between different constructs is being studied, equivalence of factor loadings (ie, metric invariance) is required. However, if the purpose is to compare subgroup means of a certain construct, be it through factor analysis or composite scores, additional equivalence of intercepts is required (ie, scalar invariance). Invariance is typically assessed based on a structural equation modelling (SEM) framework, but it can also be tested using an item response theory framework or a combination of both approaches.[13] A recent compilation of recommendations by Putnick *et al* provides more detail on the SEM framework approach.[13]

CONCLUSION

In conclusion, lack of in-person access to participants and timeliness may have pushed researchers to use online surveys and rating scales in particular. When researchers consider using this approach, they need to balance the added value of their research against the potential drawbacks such as selection bias and the use of non-validated or poorly validated scales. Moreover, state-of-the-art analysis of latent variables often requires tedious and advanced modelling techniques. While using these methods can be particularly useful during the current pandemic, authors, readers and reviewers should take a critical stance towards the results of such studies, even when sample sizes are large.

ORCID iDs
Jeroen De Man http://orcid.org/0000-0002-3196-8757
Edwin Wouters http://orcid.org/0000-0003-2268-3829

REFERENCES

1 O'Connor DB, Aggleton JP, Chakrabarti B, *et al*. Research priorities for the COVID-19 pandemic and beyond: a call to action for psychological science. *Br J Psychol* 2020;111:603–29.
2 Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of Internet E-Surveys (cherries). *J Med Internet Res* 2004;6:e34.
3 Hobart JC, Cano SJ, Zajicek JP, *et al*. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol* 2007;6:1094–105.
4 Eysenbach G, Wyatt J. Using the Internet for surveys and health research [online]. *J Med Int Res* 2002:76–94.
5 Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* 2014;7:342–6.
6 De Man J, Buffel V, van de Velde S, *et al*. Disentangling depression in Belgian higher education students amidst the first COVID-19 lockdown (April-May 2020). *Arch Public Health* 2021;79:3.
7 Ali SH, Foreman J, Capasso A, *et al*. Social media as a recruitment platform for a nationwide online survey of COVID-19 knowledge, beliefs, and practices in the United States: methodology and feasibility analysis. *BMC Med Res Methodol* 2020;20:1–11.
8 Kampen JK. Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables. *Measurement* 2019;137:428–34.
9 Revelle W. An introduction to psychometric theory with applications in R [online], 2020. Available: http://www.personality-project.org/r/book/
10 Zinbarg RE, Revelle W, Yovel I, *et al*. Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika* 2005;70:123–33.
11 Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. *Exp Aging Res* 1992;18:117–44.
12 Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ Res Methods* 2000;3:4–70.
13 Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev Rev* 2016;41:71–90.