Organisation for Economic Co-operation and Development

EDU/WKP(2022)10

Unclassified **English - Or. English**

1 June 2022

DIRECTORATE FOR EDUCATION AND SKILLS

The analytical value of non-probability samples in the context of TALIS: A review of current practices in the use of non-probability samples in comparative, cross-national research

OECD Education Working Paper No. 272

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Gabor Fülöp, Gabor.Fulop@oecd.org
Francesco Avvisati, Francesco.Avvisati@oecd.org

JT03496629

Acknowledgements

The authors thank Valerie Frey, Fernando Galindo-Rueda, Ruochen Li and Fabrice Murtin for their careful review and thoughtful feedback. The authors also wish to thank the following people for providing information about their respective survey projects: Daphne Nathalie Ahrendt and Eszter Sandor (Eurofound Living, Working and COVID-19 e-survey); Jason Ferris and Adam Winstock (Global Drug Survey 2021); Valerie Frey (OECD Risks that Matter Survey 2020), Fernando Galindo-Rueda (OECD Science Flash Survey 2020; OECD International Survey of Science 2021), Alison Kennedy (UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education), Fabrice Murtin (Trustlab survey), Mike Fisher and Vikas Pota (T4 Teachers and Technology Global Survey), Ernesto Villalba-Garcia (Cedefop: COVID-19 pandemic and career guidance systems and policy development – Joint survey by international organisations). The authors also thank Emily Groves, Lina Nguyen and Hélène Guillou for editorial support and Pablo Fraser for managerial support.

Abstract

The appeal of non-probabilistic surveys has been on the rise given the costs and decreasing response rates associated with probabilistic surveys. Yet the non-random selection of respondents into non-probabilistic surveys leads to inaccurate estimates if there are systematic differences in relation to the variable of interest between the self-selected respondents to the survey and the rest of the target population. In addition, for non-probability samples there is no general statistical theory that justifies when and why accurate inferences can be expected. This paper presents a review of established uses of non-probability samples in comparative, cross-national contexts and their value for policy. In particular, the review focuses on the rationales for using non-probability samples, the risks involved and the potential ways of mitigating these risks. The paper concludes by providing some potential roles non-probability samples could play in the context of the OECD Teaching and Learning International Survey (TALIS).

Résumé

L'intérêt pour les enquêtes non probabilistes a gagné du terrain en raison des coûts et de la baisse des taux de réponse associés aux enquêtes probabilistes. Or, s'il existe des différences systématiques en rapport avec la variable d'intérêt entre les répondants auto-sélectionnés à l'enquête et le reste de la population cible, la sélection non-aléatoire des répondants dans les enquêtes non probabilistes conduit à des estimations inexactes. De plus, pour les échantillons non probabilistes, il n'y a pas de théorie statistique générale qui justifie quand et pourquoi on peut s'attendre à des inférences valides. Ce document de travail présente une analyse des utilisations avérées des échantillons non probabilistes dans des contextes comparatifs et transnationaux et de leurs bénéfices pour les mesures politiques. En particulier, l'examen se concentre sur les justifications de l'utilisation d'échantillons non probabilistes, les risques encourus et les moyens potentiels pour atténuer ces risques. Le document conclut en envisageant certains rôles potentiels que les échantillons non probabilistes pourraient jouer dans le contexte de TALIS.

Table of contents

Tables

Figures

Boxes

1. Introduction

In the presence of unexpected events (such as the recent school closures due to COVID-19) there may be interest in complementing traditional surveys based on probability samples – such as the OECD Teaching and Learning International Survey (TALIS) – with insights from the responses of self-selected participants obtained within a much shorter time frame than a survey based on random sampling methods may permit.

Concerns about the cost and time required for probability sampling – in particular, to compile sampling frames that cover the entire target population – as well as potential non-response issues that increasingly affect probability samples, have led to growing interest and use of non-probability samples in survey research. The parallel, widespread availability of web survey tools has led some to wonder whether non-probability sampling methods (such as surveys of volunteers recruited online) might be an acceptable alternative (Baker et al., 2013[1]). While particular surveys based on non-probability samples and the methods used to analyse them have been shown to perform well in specific circumstances, such as in predicting election outcomes (Goel, Obeng and Rothschild, 2015[2]; Graefe, 2016[3]; Wang et al., 2015[4]) or providing timely estimates of the state of the labour market (Foote et al., 2021[5]), for most purposes and circumstances probability samples continue to be the recommended method (Cornesse et al., 2020[6]). In general, the extent to which estimates based on non-probability samples can be generalised beyond the respondents, and are representative of a broader population of interest, remains unknown (Baker et al., 2013[1]; Cornesse et al., 2020[6]; Tourangeau, Conrad and Couper, 2013[7]).

In this context, the TALIS Governing Board has included a project on the analytic value of non-probability samples among its research and development activities for the current biennium to answer, in particular, the following questions:

- Would non-probabilistic sampling approaches that use TALIS-based instruments increase the timeliness of TALIS insights? What are the risks involved? Is it possible to identify selection and response biases and to correct for these?
- Can non-probabilistic sampling approaches be used as part of the TALIS instrument development cycle?

This working paper presents a review of established uses of non-probability samples in cross-national contexts and their value for policy. The review aims to identify new and more agile ways of meeting the goals of TALIS, based on the experience of active OECD projects and selected non-OECD activities that rely on non-probability samples for inferences at the country level.[1] Yet it is important to highlight that the research and development (R&D) project on the analytic value of non-probability samples is not motivated by a desire to move away from the probabilistic design of the well-established and high-quality TALIS survey. Rather, the project aims to explore the potential for complementing the traditional TALIS survey. The main objectives of the review are to:

- consider whether the specific rationale for using non-probability samples instead of probability samples could apply in the context of TALIS too
- document the risks involved in the use of non-probability samples and potential ways of mitigating these in the TALIS context

[1] Self-assessment tools, such as PISA for schools, are not considered: their goal is to feed back information to participants, without generalising beyond the sample of participants.

- assess how limitations of non-probability samples are acknowledged, and how margins of uncertainty are reported in the absence of design-based inference methods.

This working paper first introduces the main characteristics and applications of non-probability samples. In particular, it explores the scenarios when non-probability samples are more likely to be fit for purpose and discusses potential ways to mitigate the risks associated with the use of non-probability samples. The paper also takes a brief look at the use of non-probability samples in non-survey contexts. Then it presents a review of established uses of non-probability samples in comparative, cross-national contexts. Namely, it makes an attempt to document the specific rationales for using non-probability samples, the risks involved in the use of non-probability samples, the potential ways of mitigating these risks, as well as to assess how limitations of non-probability samples are acknowledged and presented in cross-national surveys. The paper also draws some lessons for TALIS from current practices in the use of non-probability samples in other comparative, cross-national surveys that are included in the review. The paper concludes by providing some potential applications of non-probability samples that could be considered in the context of TALIS.

Box 1. A brief summary: potential roles non-probability samples could play in the context of TALIS

Relying on non-probabilistic sampling approaches in the context of TALIS would likely lead to a substantial loss in accuracy and the impossibility of quantifying uncertainty in findings, which makes non-probabilistic sampling unfit for achieving the purposes of the TALIS main survey. However, as the review of current practices in the use of non-probability samples in comparative, cross-national surveys shows, there can be scenarios where using a non-probability sample in the TALIS context can be considered. For instance, non-probability samples fit better for exploring relationships among a broad set of characteristics rather than precisely describing those characteristics in the population of interest. In addition, the TALIS main survey could serve as a reference data source to reduce the risks inherent in non-probability samples, to the extent possible. The potential roles non-probability samples could play in the context of TALIS include:

A complementary module with a longitudinal design exploring trends and relationships

Non-probability sampling could contribute to TALIS through a complementary module with a longitudinal design. For example, a panel of volunteer teachers or school leaders, among the sampled participants of the core TALIS survey, could participate in follow-up surveys. Collecting panel data would allow examination of the same teachers or school leaders to detect any changes in their practices, beliefs and well-being that might occur over a period of time. Exploring relationships (between baseline and follow-up levels of a same variable, and across constructs) rather than aiming to provide accurate point estimates can be more efficiently supported through non-probabilistic sampling.

Field-trial data collection

The field-trial data collection, which is implemented to validate the TALIS instruments and derived measures and to trial the operational procedures, follows a probabilistic design. However, non-probability samples could play a role in the development phase (e.g. assessing model fit and item parameters for item-response-theory models or piloting survey instruments) of the core TALIS survey. Relying on non-probabilistic sampling for the field trial could lead to more timely access to field-trial data. Yet it is important to ensure that the operational component of the field trial, whose main goal is to rehearse the sampling and survey operations ahead of the main study, is fulfilled.

An exploratory survey with scoping and outreach purposes

Moving away from generalisability, non-probabilistic sampling approaches can be used to learn about the existence of some unknown phenomenon or to gain some very preliminary understanding of a phenomenon. In the case of TALIS, such an exploratory survey based on non-probability sampling could be used to identify new topics that are deemed relevant by the teaching profession to include in the survey. Such an exploratory survey could result in the development of new questionnaire items for the main survey or new thematic modules. In addition, it could also provide a vehicle for a more direct and timely dialogue between the wider teaching profession and policy makers.

2. Main characteristics and applications of non-probability samples

Sampling refers to the methods used to enlist a study population from a wider population of interest with a view to drawing inferences about the population of interest. A probability sample is one in which every element in the population of interest has a nonzero and calculable probability of selection: no elements are omitted by design, and researchers can assign a probability of selection to each element (Tourangeau, Conrad and Couper, 2013, p. 11[7]). Non-probability samples are any type of sample that violates either of the aforementioned conditions. Thus, in non-probability samples the selected units in the sample have an unknown probability of being selected and some units of the target population may even have no chance at all of being in the sample. The selection of units is based on factors, such as convenience, prior experience or the judgement of a researcher.[2]

Non-probability sampling is a collection of methods, rather than a well-defined framework; it includes, in particular:

- **Convenience samples**: a form of non-probability sampling in which the ease with which potential participants can be located or recruited is the primary consideration. This includes volunteer samples as one example. Volunteer samples are frequently used in social science research. In some cases, volunteers sign up to join a panel for a certain period and are then asked to take part in multiple studies; this allows for collecting more information about participants. This information can be used for screening participants (to ensure that all participants are members of the population of interest) and for conducting post hoc adjustments, which aim to match the sample characteristics with the known characteristics of the target population.
- **Purposive samples**: where expert judgement is used to select participants that “represent” the diversity in the population of interest. The relevant dimensions of diversity are chosen in relation to the specific theme and focus of the research questions.
- **Quota samples**: a form of non-probability sampling in which the samples are selected based on the probability proportionate to the distribution of a variable in

[2] See also the *OECD Glossary of Statistical Terms* (OECD, 2008[38]).

the population. Quota samples involve sample stratification combined with either convenience or purposive sampling.

- **Network samples (e.g. snowballing)**: where recruitment of participants is driven by referrals from previous respondents. These methods are used for, in particular, surveys of rare populations where no (complete) sample frame for the population of interest exists and the cost of finding eligible members based on a larger sampling frame (which includes both eligible and non-eligible cases) may be prohibitive (in some conditions, network samples may also be treated as probability samples).

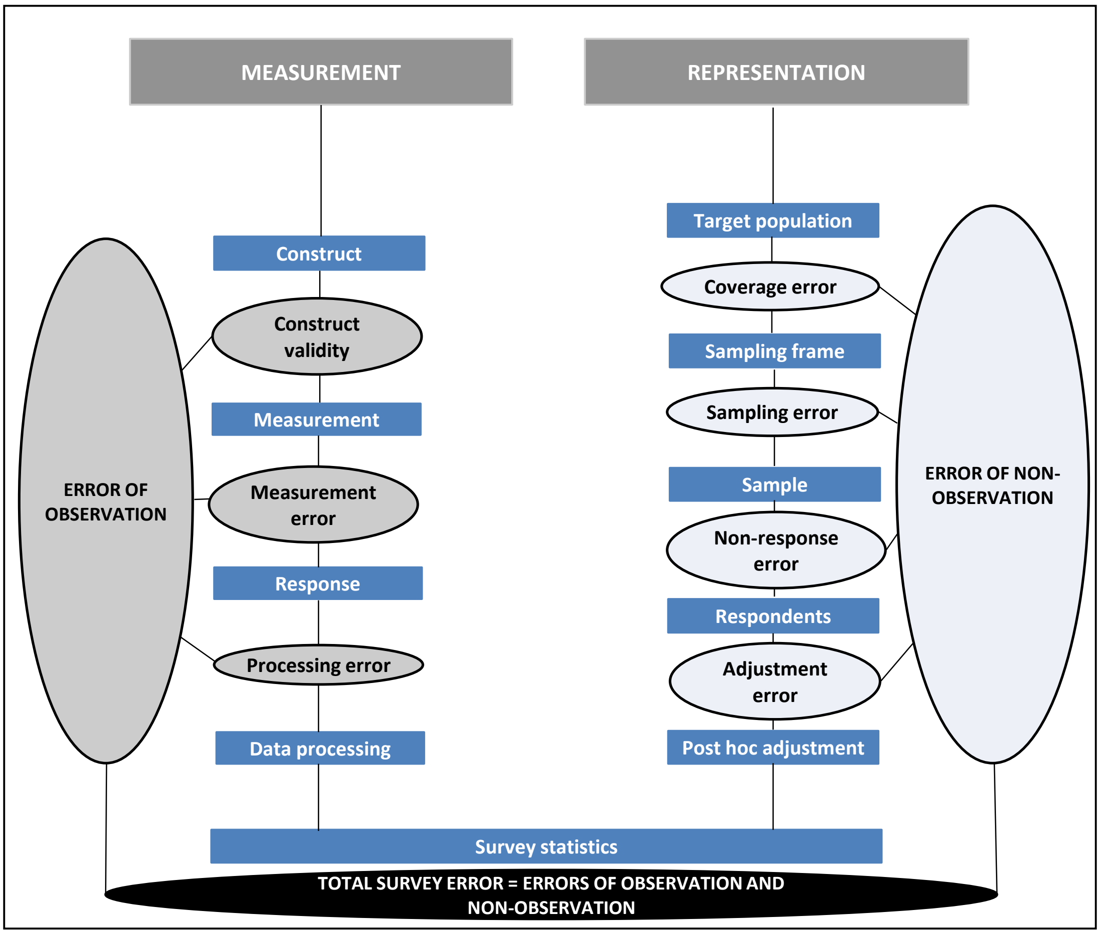
The goal of most surveys is to describe, across multiple dimensions, a population of interest. The goal of international surveys is to compare these descriptions and identify meaningful variation across countries. This goal places a strong emphasis on "accuracy", i.e. the degree to which estimates correctly measure what they are designed to measure, within an acceptable margin of error (OECD, 2011[8]).

The total survey error framework – as presented by Groves et al. (2009[9]) – takes into account the various phenomena affecting "accuracy". According to the total survey error framework, there are two main conditions needed to draw accurate inferences about the larger population of interest based on answers from respondents to a survey:

- respondents' answers must accurately describe their characteristics
- the subset of the target population that participates in the survey must have similar characteristics to the larger population of interest.

Thus, error in survey statistics can arise either due to deviations from answers given to a survey question and the underlying attributes measured (i.e. "measurement error" or "error of observation") or due to deviations of a statistic estimated on sample from that on the larger population of interest (i.e. "representation error" or "error of non-observation") (Groves et al., 2009, p. 40[9]). As shown by the total survey error framework, both observation and non-observation errors can be disentangled into various sources of errors that can lead to inaccurate survey statistics (Figure 1). Errors, both in terms of bias and variance, can arise between the different steps in the survey process. By focusing solely on representation, misalignment between the target population and the sampling frame can lead to under-coverage (or over-coverage) of the sampling frame with respect to the target population (i.e. coverage error). The sampling phase introduces error by design, since not all persons included in the sampling frame are measured (i.e. sampling error). While sampling variance refers to the variability linked to the fact that different samples can be drawn from the sampling frame, sampling bias arises when some members of the sampling frame, who would have distinctive values in relation to the survey statistics, have no chance (or reduced chance) of selection. The next potential source of error is related to the rate of actual respondents. Surveys rarely succeed in receiving responses from all people who are sampled. Non-response error occurs if the values of survey statistics estimated based only on respondents' answers differ from those based on the entire sample. In addition, the adjustment phase that aims to improve the sample estimates given the coverage, sampling and non-response errors can also have an influence on the accuracy of survey statistics.

Figure 1. Total survey error framework



Source: Adapted from Groves, R. et al. (2009[9]), *Survey Methodology, Second Edition*.

In the context of non-probabilistic surveys, the total survey error framework is useful in so far that it highlights the various potential sources of error. For instance, it draws attention to measurement errors. However, the framework does not fit well to non-probabilistic sampling approaches when it comes to describing errors related to representativeness. In the context of non-probability sampling, where often there is no sampling frame to start with, departures from a sampling frame are unknown and, as a result, errors that that can arise at each stage of the sampling process cannot be quantified (Cornesse et al., 2020[6]; Mercer et al., 2017[10]).

Non-probability samples are prone to potential bias due to uncontrolled selection. This leads to biased estimates if there are systematic differences in relation to the variable of interest between the self-selected respondents to the survey and the rest of the target population. For instance, Schaurer and Weiß (2020[11]) show selection bias for measures of personality traits among non-probabilistic online survey respondents. The authors find that the participation in an online convenience survey focusing on the outbreak of the COVID-19 pandemic in Germany is related to compliance with measures that can minimise the risk of being infected.

The selection bias inherent in non-probability samples is similar to bias due to non-response in the context of probability samples. The general decline in response rates in surveys based on probability samples raises concerns about the potential for biased results and argues for the value of non-probabilistic approaches given their larger sample sizes (Beaumont, 2020[12]). In the face of high non-response, probability samples can resemble non-probability samples. However, it is important to note that, contrary to the general trend, TALIS continues to attain high response rates – at the lower secondary level, for instance, the overall response rate of teachers is around 85% on average across TALIS participants (OECD, 2020[13]).

The growing popularity of non-probabilistic sampling is closely related to the rise in Internet access and use. Recruiting respondents on line has advantages given its speed and low costs, but it also entails risks. Non-probability samples that are administered online (e.g. in the form of open-link web surveys or opt-in online panels) not only present potential bias due to uncontrolled selection, but also due to non-coverage of large portions of the target population(s). In the case of online surveys, the bias due to non-coverage refers to the differences between the population that theoretically could answer the survey (e.g. population with Internet access) and the target population (i.e. full population) (Tourangeau, Conrad and Couper, 2013[7]). Excluding the population without Internet access from an online survey leads to biased estimates if those with Internet access differ from those without it on characteristics that are likely to be of interest in the survey.

The review of various empirical studies assessing and comparing the accuracy of estimates based on probability and non-probability samples by Cornesse et al. (2020[6]) concludes that probability surveys have significantly higher accuracy than non-probabilistic surveys even after post hoc adjustments. Research also shows that the potential for bias tends to be higher for non-probabilistic surveys than for probabilistic surveys with low response rates (Bethlehem, 2016[14]; Dutwin and Buskirk, 2017[15]). In addition, Cornesse et al. (2020[6]) cite studies – see, for instance, Brüggen, van den Brakel and Krosnick (2016[16]), Dutwin and Buskirk (2017[15]) and Macinnis et al. (2018[17]) – that disentangle the mode and sampling effects and find that both offline and online probabilistic surveys are more accurate than non-probabilistic online surveys.

The two sampling approaches also differ in the extent to which the accuracy of estimates can be tested and justified. In the case of probability sampling, the assumptions that allow for drawing general conclusions from a sample, to avoid bias and to quantify (and minimise) error, are well understood for samples drawn with known or calculable probabilities of selection (Baker et al., 2013[1]).

In contrast, the assumptions required to draw general conclusions from a non-probabilistic sample are embedded in a statistical model, which constrains the variability in the population, forcing it to follow particular rules. The “model” used to make general statements about the target population based on non-probability samples may remain implicit. For example, similarly to adjustments for non-response in the case of probability samples, when post hoc adjustments are used to claim that a non-probability sample is “representative” of a target population and can be used to quantify the prevalence of a certain behaviour, the researcher is (often implicitly) making strong assumptions about all unobserved variables and their covariance with the behaviour of interest. Such assumptions can be more easily defended (but not tested) when interest lies in a specific behaviour (such as voting intentions): in this case it may be possible to identify (*ex ante*) all the observable confounders (e.g. prior voting behaviour), to measure them reliably, and to access reliable measures of the same confounder for the target population, even when these confounders extend beyond demographics, to include attitudes – such as interest in the topic of the survey, or a reasonable proxy (Ansolabehere and Schaffner, 2014[18]).

In surveys that cover a large range of topics, the potential confounders are likely different depending on the topic. In this situation, different adjustments may be needed for different analyses – and not all of these adjustments may be possible. Indeed, comparable, external measures for all variables that may influence both participation in the sample and the behaviour of interest may not be available. This makes non-probability samples quickly unpractical for large, multi-purpose surveys, whose main goal is to “describe” the population of interest.

In contrast to probability samples, there is no general statistical theory of non-probability sampling that justifies when and why accurate inferences can be expected: validity depends on the topic and survey, and ultimately rests on untested assumptions about the characteristics that make (or do not make) the sample different from the population, and how those characteristics relate to the topic of interest (Cornesse et al., 2020, p. 7[6]). Moreover, in the lack of general statistical theory underpinning non-probability sampling, design-based inference to quantify margins of error is not possible.

2.1. Scenarios when non-probability samples are more likely to be “fit for purpose”

Despite the potential risks associated with non-probability sampling, in certain cases there might be a rationale for relying on convenience samples of volunteers or other forms of non-probability samples.

2.1.1. Timeliness, feasibility and value-for-money

Criteria other than accuracy might be considered when evaluating if survey data are “fit for purpose”. These criteria include: relevance, timeliness, accessibility, interpretability and consistency (Baker et al., 2013[1]; OECD, 2011[8]). In particular situations, non-probability samples may be superior to probability samples in terms of timeliness. In addition, feasibility and cost (value-for-money) is also an important factor that often motivates the use of non-probability samples; in some situations, the cost of sampling and of convincing non-volunteers to participate may be prohibitive and justify the use of non-probability samples. In such cases, the alternative to a non-probability sample is a highly imperfect probability sample. As discussed previously, the representativeness of probability samples and, hence, the chances of providing accurate estimates, decreases as non-sampling errors related to measurement, coverage or non-response become substantial.

2.1.2. Sample selection process is unrelated to the phenomenon of interest

In some cases, researchers may claim that the sample selection process is unrelated to the phenomenon of interest. For example, researchers who are interested in universal physiological or psychological phenomena may presume that they can find evidence of such phenomena in any sample, irrespective of how it has been selected. For instance, in medical research, it is not uncommon to use samples of volunteers to make general claims about the effectiveness of a new vaccine or medicine in a particular age group. The same argument is also often invoked to justify the “external validity” of results obtained in randomised control trials. As shown by these examples, non-probability samples are used in many areas of research, including to inform policy.[3]

[3] Also note that not all inferences and generalisations from international surveys are and can be based on the sampling design. In particular, countries (or other jurisdictions) that participate in international surveys are not selected based on a probabilistic sampling process; instead, they are self-selected according to a non-probabilistic process. Thus, any analysis based on country-level associations that is generalised beyond the sample (even to the same countries, at a different point in time) cannot rely on the properties and tools developed for probabilistic samples.

2.1.3. Exploring relationships rather than point estimates

Sometimes the justification is empirical, rather than theoretical; prior research has found that, even when means and proportions in non-probability samples were biased, relations among variables remained similar across probability and non-probability samples (Pasek, 2015[19]). Indeed, research that focuses more on exploring relationships among a broad set of characteristics rather than precisely describing those characteristics in the population of interest tends to rely more often on non-probability samples (Baker et al., 2013[1]); the difficulty is that such an empirical justification may not extend to all surveys and topics. Pilots and field trials aiming at instrument development – such as analyses of factor structure or ranking of question quality – within the development of probabilistic surveys illustrate well how non-probability samples can be useful in certain contexts.[4]

2.1.4. Focus is on a narrow set of estimates (e.g. electoral polling)

As mentioned previously, non-probability samples can also perform well in identifying parameters of the population distribution in specific situations (e.g. electoral polling): however, they are less suitable in the context of more complex surveys that measure many different phenomena (Baker et al., 2013[1]). In particular, non-probability samples can work well in the context of electoral polling since these surveys are designed to yield only a handful of estimates on a related set of outcomes that require the control of only a small set of covariates (Baker et al., 2013, p. 103[1]). In addition, the availability of external benchmarks (such as election results in the context of electoral polling) can also help model development. However, most surveys aim to produce many estimates across different topics and domains, requiring a larger set of covariates, which makes non-probability samples less useful as a general approach to data collection.

2.2. Ways to mitigate the risks associated with the use of non-probability samples

The literature proposes various ways to reduce risks associated with non-probability samples. These include: the collection and use of variables that are correlated with both inclusion in the sample and the outcome of interest; certain control over the sampling process; and post hoc adjustments, similar to the ones used in the context of probability samples.

2.2.1. Identifying theoretically grounded confounders as a starting point

The availability and use of good confounders – variables that are correlated with both inclusion in the sample and the outcome of interest – initially for the research design, and later on at the sampling and adjustment phases, is important for reducing the risks associated with selection bias (Mercer et al., 2017[10]). For example, in the case of electoral polling, many outcome variables of interest will be related to respondents' underlying political engagement and partisanship. Ensuring that these confounders are available can greatly reduce risks of non-coverage and self-selection and provide information about the generalisability of the survey statistics.

Based on the framework proposed by Mercer et al. (2017[10]) it can be determined whether or not self-selection could lead to biased results by exploring the following issues:

[4] Examples for such pilot studies and field trials include the Assessment of Learning Outcomes in Higher Education (AHELO) feasibility study, the Innovative Teaching for Effective Learning (ITEL) Teacher Knowledge Survey pilot study and the Programme for International Student Assessment (PISA) field trials.

- Exchangeability: Are all confounding variables known and measured for all sampled units?
- Positivity: Does the sample include all of the necessary kinds of units in the target population, or are certain groups with distinct characteristics systematically missing?
- Composition: Does the sample distribution match the target population with respect to the confounding variables, or can it be adjusted to match?

In practice, the above framework calls for the following elements at the sample design phase to avoid selection bias:

- Exchangeability: Selecting the right combination of confounders for use in quotas, weighting or other modelling.
- Positivity: Selecting a data collection protocol that is capable of reaching in sufficient numbers all necessary kinds of units in the target population.
- Composition: Availability of distributional information for quota targets, raking parameters, post-strata for multilevel regression and post-stratification (MRP).

Thus, achieving exchangeability requires the right combination of confounders. Yet it has to be noted that, even with strong theoretically grounded confounders, achieving exchangeability is usually very challenging (Mercer et al., 2017[10]).

2.2.2. Importance of minimum levels of control over the sampling process

Some control over the sample selection is essential to improving the accuracy of inferences from non-probability samples (Baker et al., 2013[1]). With respect to surveys whose main goal is to assess opinions, the dangers of non-probability samples are particularly acute in the case of open-link web surveys, and when the survey is used in order to inform decisions that may affect the survey respondents themselves. In such surveys, three additional risks must be considered (Bethlehem, 2017[20]):

- Can people outside the target population participate in the survey? (Screening for eligibility can be more difficult in open-link web surveys.)
- Can participants respond more than once (on the same, or on a different computer)?
- Is it possible for a group of people to manipulate the outcomes of the open-link web survey?

Although the risks of non-genuine responses are context dependent and, in many cases, can be considered minimal, they cannot be fully excluded either. While controlling these risks does not necessarily require a probability sample, it requires some form of sampling frame or authentication mechanism.

2.2.3. Applying post hoc adjustments to reduce coverage and selection bias

As already mentioned in the context of the total error framework, even in the case of probability samples there are post-survey adjustments that aim to improve the accuracy of survey statistics in the face of errors of non-observation (e.g. coverage, sampling and non-response errors) (Groves et al., 2009[9]). These post hoc adjustments typically rely on information about the target or frame population, or response rates to adjust for systematic biases in probability samples. Similar adjustment procedures are used in the case of non-probability samples to correct for potential bias due to non-coverage and uncontrolled selection. These methods apply weights to make the sample of respondents resemble the target population more closely (Tourangeau, Conrad and Couper, 2013[7]). They include the following:

- **Post-stratification**: This method adjusts the sample weights in such a way that the sum of weighted samples equals the known population total within each mutually exclusive group of the target population (i.e. adjustment cell). This is a relatively easy method that can eliminate the bias due to non-coverage and uncontrolled selection provided that, within each adjustment cell, the probability of a respondent completing the survey is unrelated to his or her values on the variables of interest of the survey (i.e. missing at random assumption). Post-stratification reduces bias (without fully eliminating it) as long as the covariance between the probability of participation and the survey variable is reduced after the adjusted sample weights are taken into account (Tourangeau, Conrad and Couper, 2013[7]).
- **Raking**: Similarly to post-stratification, this method also adjusts the sample weights in order to align the sample more closely with the target population. However, raking does not align the sample to the cell totals of the known population, but to the marginal totals. This means that the adjusted sample weights add up to the population total for each auxiliary variable separately (e.g. gender and education), but not necessarily to their combinations (e.g. males with college degrees or females without college degrees). Raking might be preferred to post-stratification in case the population figures are either not available, or there are only very few participants (or no participants at all) in a given adjustment cell. Therefore, raking can be a solution if the number of auxiliary variables used in the weighting is too high for a cell-by-cell adjustment to be feasible. Raking reduces or eliminates bias under the same conditions as post-stratification (Tourangeau, Conrad and Couper, 2013, p. 26[7]).
- **Generalised regression (GREG) modelling**: This method aims at improving sample estimates by assuming a linear relationship between the variable of interest and the available auxiliary variables. Unlike post-stratification and raking, GREG weighting can also easily handle non-categorical auxiliary variables. Most importantly, it can cope with cells that have no respondents. Otherwise, GREG weighting reduces or eliminates bias under the same conditions as post-stratification and raking (Tourangeau, Conrad and Couper, 2013[7]).
- **Propensity scoring**: A propensity score is the predicted probability that a respondent belongs to a certain group - for example, the probability that someone will be among those that have Internet access (versus not having access) – given a set of auxiliary variables (Tourangeau, Conrad and Couper, 2013[7]). Propensity models require a calibration or reference survey with little or no bias due to non-coverage or uncontrolled selection. Similarly to GREG weighting, propensity scoring has potential advantages over the more simple methods of post-stratification and raking in terms of flexibility and, in particular, in handling non-categorical auxiliary variables and empty cells.

As briefly presented above, all methods are motivated by some form of a "selection on observables" assumption. It is assumed that the "…selection mechanism of the non-probability sample can be ignored conditional on the variables used in the adjustment method." (Cornesse et al., 2020, p. 12[6]).

Based on studies examining the effectiveness of the different post hoc adjustment procedures in the context of web surveys, Tourangeau, Conrad and Couper (2013, p. 35[7]) conclude that "…regardless of the exact method used, the adjustment procedures typically remove less than half of the bias in the estimates and often substantial biases remain after adjustment. Sometimes the adjustments backfire and increase the bias. Even when they reduce bias, the adjustments often come with a penalty of increased variance." The review of empirical studies assessing the accuracy of non-probability samples by Cornesse et al. (2020[6]) also shows that post hoc adjustments are not a panacea for addressing inaccuracy.

2.2.4. Transparency about the methods and assumptions applied

A clear description of the non-probability sampling method and the modelling assumptions used is essential for understanding the validity of non-probability survey estimates (Baker et al., 2013[1]; Cornesse et al., 2020[6]). Assessing the risks associated with the use of non-probability survey estimates is not simple. Each non-probability sampling method has different approaches to sampling and estimation, with different statistical properties, empirical performance and limitations. Similarly, the disclosure of key modelling assumptions and, to the extent possible, an evaluation of how deviations from those assumptions affect the accuracy of estimates are important for assessing the quality of the estimates (Baker et al., 2013[1]).

2.3. Non-probability samples in non-survey contexts

Non-probability samples are also regularly used in the context of self-assessment tools.[5] These tools allow respondents to compare themselves to a well-established benchmark (e.g. national average). In the case of a self-assessment, the representativeness of respondents is not an issue as long as the primary objective is to provide information to the respondent rather than to generalise responses at a more aggregate level. Yet as soon as responses from self-assessment tools are used to make inferences about the general population, the risks and challenges associated with the use of non-probability samples in a survey context arise. However, in the interests of space, a full consideration of self-assessment tools relying on non-probability samples are not discussed in detail within this report.

Statistical analyses that harness "big data" face analytical challenges similar to those based on surveys administered to non-probabilistic samples. In such cases, large volumes of data collected from personal interactions with digital platforms, through web scraping, or by activity trackers, are used in order to describe a static or evolving phenomenon, or as input for model-building. Although the data are not collected through a survey instrument, the analytical challenges associated with drawing general conclusions are similar to those associated with non-probability samples. As a result, the techniques used in the analysis of non-probabilistic surveys can be expected to benefit from advances in the field of "big data" analysis. It will be worthwhile monitoring the developments in big data analysis seeing it is a field only likely to grow considerably and with potentially wide-reaching ramifications.

[5] An example for such a self-assessment tool is a specific application attached to the OECD Better Life Index, in which respondents can create and compare their own composite index based on country-level indicators that are based on administrative data and probabilistic surveys. Another example is the Education & Skills Online Assessment, which is an assessment tool designed to provide individual-level results that are linked and comparable to the OECD Survey of Adult Skills (PIAAC) measures of literacy, numeracy and problem solving in technology-rich environments. Other examples for self-assessment tools include a customisable tool that helps schools assess where they stand with learning in the digital age (i.e. Self-reflection on Effective Learning by Fostering the use of Innovative Educational technologies [SELFIE]) and also online personality tests that provide respondents immediate scoring and feedback regarding their personality (e.g. www.outofservice.com).

3. A review of current practices in the use of non-probability samples in comparative, cross-national surveys

This section provides a review of established uses of non-probability samples in the context of cross-national surveys. More specifically, it reviews the survey projects' specific rationale for using non-probability samples instead of probability samples and documents their experience in terms of the main advantages and challenges of relying on a non-probabilistic design. In addition, it also reviews the selected survey projects from the viewpoint of the risks involved in the use of non-probability samples, the potential ways of mitigating these risks, as well as how the limitations of non-probability samples are acknowledged and presented. All the survey projects included were reviewed based on the same set of questions (Table A.1).

The review focuses on cross-national surveys that rely on non-probability samples and potentially aim to make inferences at the country level. Various OECD and non-OECD projects that rely on non-probability samples are identified and included in the review (Table A.2). A snowballing technique was applied to identify cross-national surveys that apply non-probabilistic sampling. Authors of already identified surveys were consulted about the existence of other cross-national non-probabilistic surveys.

In total, the review includes 11 surveys (Table A.2). Although surveys targeting teachers are of particular interest for this review, other cross-national surveys relying on non-probability samples that have a different focus and context are also included. Only 3 out of the 11 projects surveyed teachers (i.e. Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19, T4 Teachers and Technology Global Survey, UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education). Two out of these three surveys targeting teachers, the Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19 and the T4 Teachers and Technology Global Survey, were related to the challenges and disruption that educations systems faced in the wake of the COVID-19 pandemic. The Joint OECD-Harvard Graduate School of Education survey aimed to assist education leaders in their efforts to maintain education continuity amidst this pandemic (OECD, 2020[21]). It collected information from a wide variety of stakeholders, from teachers and school principals to senior government officials and education administrators. The focus of the T4 Teachers and Technology Global Survey was to assess the impact of the pandemic on teachers and learners globally, in particular by focusing on issues around technology use and digital resources (Pota et al., 2021[22]). The third survey targeting teachers, the UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education, looked at teachers' feelings of readiness to teach themes related to "learning to live sustainably" and "learning to live together in peace" (Sustainable Development Goal [SDG] Target 4.7) (UNESCO; Education International, 2021[23]).

The other eight surveys included in the review cover a wide variety of themes and contexts (Table A.2). There are surveys looking into expectations about the government and its policies (OECD, 2021[24]); studying social preferences, generalised trust and trust in institutions using experimental games (Murtin et al., 2018[25]); providing a snapshot of career guidance delivery, services, usage and careers learning during the pandemic (Cedefop; European Commission; ETF; ICCDPP; ILO; OECD; UNESCO, 2020[26]); looking at science policies and the scientist community's perceptions (i.e. OECD Science Flash Survey 2020 and OECD International Survey of Science [ISSA] 2021); capturing the economic and social effects (i.e. impact on well-being, health and safety, work and telework, people's work-life balance and financial situation) of the COVID-19 pandemic in the European Union (EU) (Eurofound, 2021[27]; Eurofound, 2020[28]); providing evidence on how lesbian, gay, bisexual, trans or intersex (LGBTI) people experience their human and fundamental rights (European Union Agency for Fundamental Rights, 2020[29]); and attempting to identify new drug trends and drug-related harm (Winstock et al., 2021[30]).

It is important to note that self-assessment tools and instrument-development pilots that are part of the preparations for a probabilistic survey (e.g. field trial) were, as aforementioned, considered out of scope and, hence, excluded from the review.

The surveys reviewed cover various non-probabilistic sampling methods, such as quota and convenience samples, as well as a combination of convenience, purposive and snowballing samples (Table A.2). The surveys relying on quota samples, such as the OECD Risks that Matter Survey 2020 and the Trustlab survey, can be considered nationally representative according to certain demographic and socio-economic characteristics (e.g. age, gender, education level). Yet the rest of the surveys covered, which are based on either convenience samples or a combination of convenience, purposive and snowballing samples, do no claim to be representative of a well-defined population.

Except for the surveys based on quota samples, where at least 1 000 responses are collected per country, the number of respondents and countries covered by the rest of the surveys tend to vary considerably both within and across survey projects (Table A.2).

While the majority of the surveys included in the review make inferences at the country level and comparisons across countries, this is not always the case (Table A.2). There are surveys that refrain from making inferences at the country level and only report about the population of respondents (i.e. Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19; Cedefop: COVID-19 pandemic and career guidance systems and policy development – Joint survey by international organisations, OECD Science Flash Survey 2020, UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education).

3.1. Rationales for using non-probability samples

Across the surveys reviewed, timeliness and feasibility were often mentioned as rationales for using non-probability samples. During the disruption caused by the COVID-19 pandemic, standard survey operations were on hold, while there was also a need to fill data gaps in an agile way. In this context, an online survey was practically the only feasible option to survey people (i.e. Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19, Cedefop: COVID-19 pandemic and career guidance systems and policy development – Joint survey by international organisations, OECD Science Flash Survey, T4 Teachers and Technology Global Survey, Eurofound Living, Working and COVID-19 e-survey). In addition to being agile and providing information in a timely manner, surveys based on non-probability samples can also serve as communication tools that help in raising awareness of certain issues and topics as well as in promoting other work done by the sponsor organisation.

Cost efficiency is the other main motivation for using non-probability samples across various surveys reviewed. As probability samples require a more complex statistical infrastructure, their costs are considerably higher. Costs were a major rationale in the case of the surveys relying on quota samples, such as the OECD Risks that Matter Survey and the Trustlab survey. These surveys might have considered applying probabilistic sampling if it was not for the costs and complexity of collecting a probability sample. Researchers for the OECD Risks that Matter Survey also raised concerns about high non-response rates in probability samples, which limit the accuracy of these samples. Yet it is not only surveys relying on quota samples that indicate costs as the main rationale for using non-probability samples instead of probabilistic sampling. For instance, the UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education that relied on a convenience sample also reported costs as the main rationale in opting for non-probabilistic sampling.

In addition, the use of non-probability sampling can be justified if the alternative is a rather imperfect probabilistic sample. For instance, in previous cycles of the OECD ISSA prior to 2021, which collected information on scientific authors, the probability samples were imperfect due to likely non-coverage of various groups of the target population (e.g. junior authors, non-academic researchers) and due to capacity restrictions on targeted large-scale mailing from the OECD. Thus, even in a probabilistic setting, post hoc adjustments after data collection were required to mitigate risks related to potential bias and results had to be framed in the context of a more selected sub-population. Thus, it was decided to switch to a non-probabilistic sampling approach that could eventually lead to a more comprehensive reach as well as larger samples sizes.

Similarly, surveys targeting "hard-to-reach" populations, such as the European Union Agency for Fundamental Rights' 2019 LGBTI survey (EU-LGBTI II Survey) and the Global Drug Survey 2021 (European Union Agency for Fundamental Rights, 2020[29]; Winstock et al., 2021[30]), opted for non-probabilistic data collection as a traditional probabilistic design would not only be very challenging to achieve but it would also be highly imperfect. For these surveys that target lesbian, gay, bisexual, trans or intersex (LGBTI) people or those who use drugs (both licit and illicit), it is very challenging at best to achieve a representative random sample in the absence of sampling frames and reliable, detailed information about the target population in terms of its size, characteristics and composition (European Union Agency for Fundamental Rights, 2020[31]). Relying on probability sampling methods is particularly challenging (i.e. costly) when the prevalence of the target population is low. And, even if random sampling could be pursued, the rate of non-response among LGBTI people and drug users would be especially high due to social stigma and even prosecution depending on the country. The anonymity and confidentiality, which is crucial to engaging these target populations, is considered to be better ensured by open-link web surveys using an encrypted online platform than traditional surveys conducted face to face or by telephone.

Beyond feasibility, for the Global Drug Survey project using non-probability samples was also motivated by the fact that the survey aims to explore relationships and trends rather than focus on point estimates. As argued by the Global Drug Survey project, having an over-representation of people who have greater interest in or experience with drugs is not a concern, as their interest is in the early identification of new drug trends and drug using behaviours rather estimating prevalence of drug use in the community.[6]

For most surveys reviewed there are no similar surveys based on probability samples to which they could link the non-probabilistic samples. However, in those few cases where identical or comparable questions were administered through probability samples, results are reported to be roughly aligned with similar surveys based on probability samples (e.g. OECD Risks that Matter Survey, Eurofound Living, Working and COVID-19 e-survey). The online survey by Eurofound adopted instruments from their well-established surveys that are based on probability samples: the European Quality of Life Survey (EQLS) and European Working Conditions Survey (EWCS). While the Eurofound Living, Working

[6] As Winstock and Ferris (2020[37]) show, based on data from the Global Drug Survey, questions about associations around drug use can be answered using data from non-probability samples. Looking at those who use nitrous oxide, the authors are able to establish the dose-response relationship between number of bulbs/balloons that are consumed and nerve damage.

and COVID-19 e-survey emphasised the usefulness of their online survey in filling data gaps, the survey project also stressed that there was no intention to replace their high-quality probabilistic surveys. They regard non-probabilistic online surveys as a complementary tool.

For the same reasons they opted for non-probability sampling (i.e. feasibility, timeliness and cost efficiency), none of the reviewed surveys expressed the intention to administer the same survey with a probabilistic sampling design.

Overall, the reviewed survey projects perceive that the rationale for using non-probability sampling is justified in their specific context, with feasibility, timeliness and cost efficiency listed as the main advantages of using non-probability sample. As cited by the Eurofound Living, Working and COVID-19 e-survey, beyond collecting timely data during the pandemic when there was an information gap, they also managed to achieve large enough sample sizes to make comparisons between different groups of respondents (with most EU member countries having over 1 000 respondents). The Global Drug Survey also highlights the advantage of collecting larger samples, using their non-probabilistic design, than the representative household surveys achieve. Another advantage of relying on non-probability sampling, as reported by the Eurofound Living, Working and COVID-19 e-survey, was to implement innovative elements in the survey, such as a panel component that allows for tracking of the evolution of the same respondents over time.

As far as the main challenges of non-probability sampling are concerned, all the reviewed projects acknowledge that potential bias due to non-coverage and uncontrolled selection may negatively affect accuracy. As highlighted by the Global Drug Survey, due to their non-random, opportunistic sampling method, their data should not be used to provide point estimates of the prevalence of drug use in the population or other generalisations about the overall population. Although there are survey projects that report a certain degree of difficulty in convincing countries to participate in a survey based on a non-probability sample due to accuracy and credibility concerns, in general, credibility and reputation are not perceived as the main challenges in opting for non-probability sampling.

Although not specific to non-probability samples, it is important to stress that sampling is only one potential source of total survey error (Groves et al., 2009[9]). It is just as important to minimise the errors related to the constructs and their measurement. As highlighted by the Eurofound Living, Working and COVID-19 e-survey, a questionnaire design characterised by careful development process, the use of already tested instruments and thorough, high-quality translation procedures are important pre-conditions of the ultimate strength and analytical value of any survey, irrespective of the sampling approach.

3.2. Risks associated with the use of non-probability samples and associated mitigation strategies at various stages of the data cycle

3.2.1. Study design

All reviewed surveys report that they were aware at the study design phase that non-probability samples could be subject to potential bias, in particular due to non-coverage and uncontrolled selection. A common source of potential non-coverage mentioned across various surveys reviewed is access to the Internet, digital literacy and use of social media, in the case of surveys conducted on line, and also the availability of the questionnaire in the language(s) of the target population. Other sources of bias due to non-coverage that were mentioned depend on the specific survey context, and include the difficulty of reaching lower-educated, elderly segments of the population, as well as those living in remote areas (e.g. the OECD Risks that Matter Survey and Eurofound Living, working and COVID-19 e-survey). In the case of the Trustlab survey, where respondents received monetary incentives for participation, high-income groups were the most difficult to reach.

Even though it is not possible to fully correct for the potential bias inherent in non-probability samples, one way to mitigate the risks associated with self-selection is to include as many variables as possible that are most likely correlated with both inclusion in the sample and the outcome of interest (i.e. theoretically grounded confounding variables) already at the study design phase. This allows to collect data for these confounding variables and to use them later on in the adjustment phases (Mercer et al., 2017[10]).

Most of the surveys reviewed made an effort to collect information on demographic variables. Yet there were fewer projects that thought ahead and also collected further potential confounding variables. The OECD Risks that Matter Survey 2020, the Trustlab survey and the Global Drug Survey 2021 all collected information on an extended range of individual characteristics. In addition to using it for participation authentication purposes, the OECD ISSA 2021 relies on the Open Register and Contribution ID (ORCID), which is a worldwide online register of researchers, to gather information on gender, age, country and publication history that can be regarded as confounders in the case of that survey. The EU-LGBTI II Survey gathered information about the respondents' affiliation with LGBTI organisations and their participation in other LGBTI surveys (including the 2012 cycle of the same FRA LGBTI survey) on the basis that these could be confounders. Nevertheless, survey projects acknowledged the risks related to not having data for all variables that could be identified as potential confounders.

Surveys administered through open-link web surveys are particularly vulnerable to potential bias in case certain risks – such as the risk that people outside the target population participate in the survey, the risk that the same participant responds more than once, and the risk that a group of people to manipulate the outcomes of the survey – are present but remain unaddressed.

Various open-link web surveys reported having relied on screening questions to address the risks associated with people outside the target population participating in the survey (e.g. UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education, Eurofound Living, Working and COVID-19 e-survey, the EU-LGBTI II Survey, Global Drug Survey). The OECD ISSA 2021 offers respondents the option of authenticating through the Open Register and Contribution ID (ORCID), which is a worldwide online register of researchers. Nearly three-quarters of participants have used this option. For the remaining quarter, e-mail authentication was required and it is possible to conduct analysis on whether those participants have a track record in research publishing. The Eurofound Living, Working and COVID-19 e-survey, which targeted people aged 18 or above, coded the compulsory question on age so that entering age under 18 ended the survey with a warning message.

Most of the online open-link surveys reviewed consider that the chances of certain participants responding more than once to manipulate the outcome of the survey on purpose, as well as other non-genuine responses, are negligible given the time required to complete the surveys, which offer no material incentives and have no high stake implications (e.g. UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education, EU-LGBTI II Survey, Global Drug Survey). Yet there are also examples of attempting to ensure single entry by each respondent. For example, the Global Drug Survey removes duplicate observations at the data processing stage (Barratt et al., 2017[32]).[7] In the meantime, some survey projects also mentioned obstacles when it came to filtering out multiple responses from the same respondent, either for practical reasons (e.g. groups of teachers with limited computer access using the same computer for filling out the survey, in the case of the UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education) or due to data protection concerns (e.g. no access to web cookies from the survey website, in the case of the Eurofound Living, Working and COVID-19 e-survey).

Overall, for all surveys administered through open-link web surveys that are included in this review, the aforementioned risks related to uncontrolled sampling are present and any resulting bias cannot be excluded.

3.2.2. Data collection

The reviewed survey projects based on quota samples (i.e. the OECD Risks that Matter Survey and the Trustlab survey) ensure that their sample is representative along the quotas, such as gender, age group education level, income level and, in the case of the OECD Risks that Matter Survey, worker status. These surveys based on quota sampling, which were implemented by private polling companies after a procurement process, tend to have at least 1 000 respondents per country.

Most of the other surveys that are based on either convenience samples or a combination of convenience, purposive and snowballing samples also made efforts to have certain control over the sampling phase. Several surveys relied on targeted recruitment strategies, such as social media advertisements (e.g. OECD ISSA 2021, T4 Teachers and Technology Global Survey, Eurofound Living, Working and COVID-19 e-survey, Global Drug Survey) and, in some cases, targeted emails as well (e.g. OECD ISSA 2021, in particular to senior management of universities and research organisations), in order to increase sample size and also to have certain control over the composition of the sample. Some of the survey projects regularly monitored the number and composition of respondents during the data collection period to collect more data from hard-to-reach segments of their target population (e.g. UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education, Eurofound Living, Working and COVID-19 e-survey). For instance, the Eurofound Living, Working and COVID-19 e-survey tried to address non-coverage by monitoring and adjusting its social media advertisements on a weekly basis by countries and/or specific sub-groups (based on age, gender and education level) that required specific targeting.

These surveys based on either convenience samples or a combination of convenience, purposive and snowballing samples achieved different sample sizes and compositions, depending on their recruitment strategies but also on the characteristics of their target populations. The Eurofound Living, Working and COVID-19 e-survey, the Global Drug Survey 2021 and the EU-LGBTI II Survey managed to collect relatively large and balanced samples across the countries covered (Table A.2). For instance, the EU-LGBTI II Survey, attained almost 140 000 responses across 30 countries and as a result surpassed its target

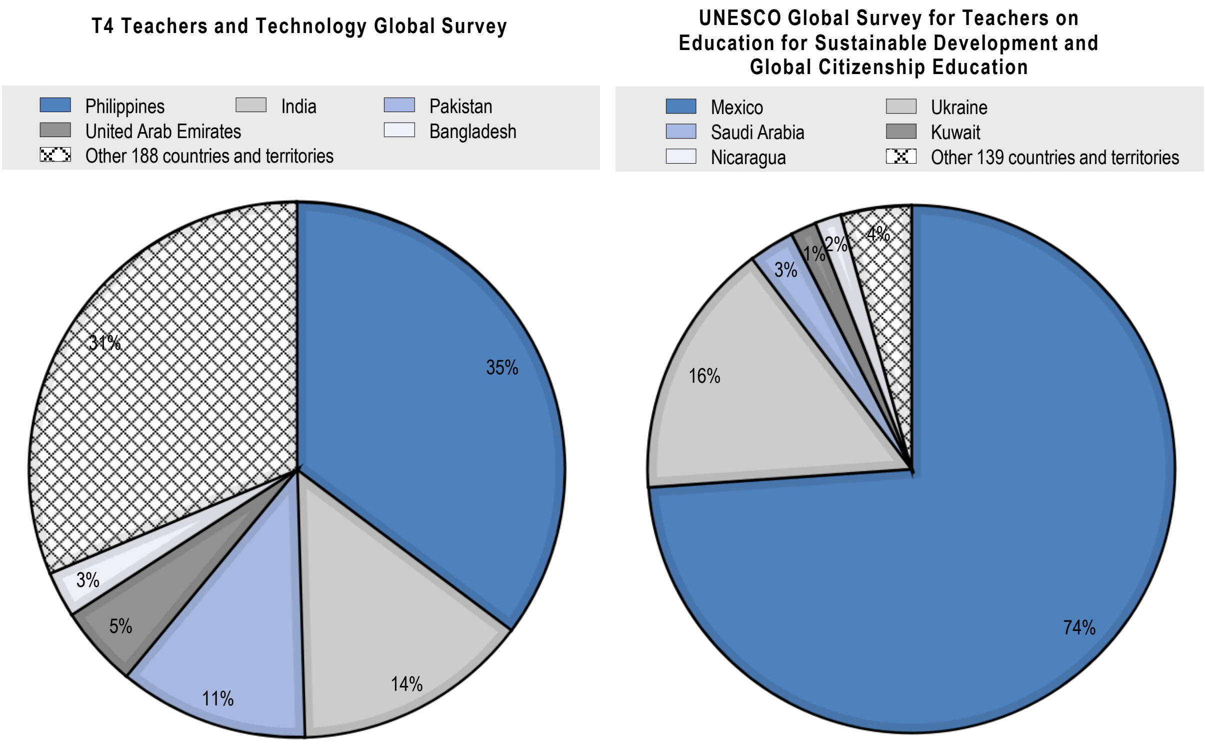
[7] Duplicate removal is two-staged. The first stage removes all records that are complete matches. The second stage removes all records where duplication is present on a series of demographic variables and drug use variables captured in the drug screen module. Demographic variables include the following: age, sex, ethnicity, educational attainment, employment type, income, country, height, weight, clubbing activity, exercise, and standard of living. Drug screen variables include the following: ever, past year, past month, age of first use, and frequency of use. Second and subsequent duplicate records are removed (Barratt et al., 2017, p. 6[32]).

sample size based on the LGBTI category and age group in almost all countries (European Union Agency for Fundamental Rights, 2020[31]).

The T4 Teachers and Technology Global Survey, which applied a combination of convenience, purposive and snowballing sampling, collected almost 21 000 responses across 165 countries (Table A.2). The UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education, which conducted a convenience sample in the form of an online open-link questionnaire, managed to reach a sample of around 58 000 responses across 144 countries. Both samples turned out to be skewed towards certain countries (Figure 2).

Figure 2. Online open-link surveys implemented in a cross-national context can be heavily skewed towards certain countries

Percentage of respondents by countries and territories



Note: Percentages based on raw counts. In order to mitigate the effects of the imbalance in response rates across countries on their own analyses, weights based on worldwide teacher population were computed by the survey teams.

Source: Adapted from Pota, V. et al. (2021[22]), *T4 Turning to Technology: A Global Survey of Teachers' Responses to the Covid-19 Pandemic*, https://t4.education/t4-insights/turning-to-technology?hsLang=en (accessed on 23 March 2022); UNESCO and Education International (2021[23]), *Teachers Have Their Say: Motivation, Skills and Opportunities to Teach Education for Sustainable Development and Global Citizenship*, https://unesdoc.unesco.org/ark:/48223/pf0000379914 (accessed on 23 March 2022).

The surveys that relied mainly on their networks and institutional partners to reach their target population, such as the Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19 and the Cedefop: COVID-19 pandemic and career guidance systems and policy development – Joint survey by international organisations, were among those with the lowest sample sizes (i.e. 1 370 and 963 respondents across 59 and 93 countries, respectively) (Table A.2). In the case of both these surveys, data collection was purposively targeting respondents from a wide range of stakeholders related to the topic that were most likely to yield appropriate and useful information on the survey topic.

3.2.3. Data processing

Among the surveys reviewed, the attention given to data processing steps, in particular to weighting, is related to the sampling method. The surveys based on quota samples, such as the OECD Risks that Matter Survey 2020 and the Trustlab survey, focused on reaching nationally representative samples along their respective quotas.

Most of the surveys using either convenience samples or a combination of convenience, purposive and snowballing samples relied on post-hoc adjustment methods to reduce the potential sampling error due non-coverage and uncontrolled selection. All the surveys reviewed that attempted to adjust the country-level estimates applied post-stratification. This technique refers to a weighting strategy that draws on known population characteristics to adjust the sample weights in such a way that the sum of weighted samples equal the known population total within each mutually exclusive group of the target population. The survey projects that applied post-stratification and provided country-level estimates included the OECD ISSA 2021,[8] Eurofound Living, Working and COVID-19 e-survey[9] and the EU-LGBTI II Survey.[10]

Some of the survey projects reviewed that did not make inferences at the country level relied on weights of geographical distribution of the target population to correct for the under- or over-representation of some countries/regions compared to their share in the global target population (i.e. T4 Teachers and Technology Global Survey, UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education).

In contrast, the Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19 aimed for equal representation of all countries in the reported estimates for the overall population of respondents. Therefore, to provide all countries the same weight in the analysis, the data that were collected from the various stakeholders were weighted by a factor equal to one over the number of respondents per country (OECD, 2020, p. 12[21]). Similarly, in the Cedefop: COVID-19 pandemic and career guidance systems and policy development – Joint survey by international organisations, all countries contribute equally to the overall figures presented in the report.

Even though most of the reviewed surveys relied on some sort of post hoc adjustment to try to reduce the potential bias inherent in non-probability samples, it is important to highlight that these adjustments (e.g. post-stratification) can only eliminate bias if the probability of a respondent completing the survey is unrelated to his or her values on the variables of interest of the survey (i.e. missing at random assumption). Re-weighting cannot adjust for those who do not have a chance to respond to the survey. Unless the missing at random assumption is fulfilled, post hoc adjustment in itself is unlikely to lead to representativeness. What is more, as shown by research, post hoc adjustment can even

[8] At the time of writing this paper, the analysis is in progress. Weights are being computed based on existing bibliographic data obtained from researchers' publication history.

[9] Weights were computed based on age, gender, urbanisation level and education level.

[10] Weights were computed based on age, affiliation with LGBTI organisations and participation in other LGBTI surveys (including the previous 2012 cycle of the same LGBTI survey). This was done to correct for possible over-representation of respondents closely affiliated with LGBTI organisations and with a higher propensity of participating in LGBTI surveys (European Union Agency for Fundamental Rights, 2020, p. 8[29]).

increase the bias or come with a penalty of increased variance (Tourangeau, Conrad and Couper, 2013[7]).

3.2.4. Reporting

All reviewed surveys acknowledge the limitations of non-probability samples when it comes to reporting results. Yet, beyond having a general note about the limitations and risks associated with the use of non-probability sampling, the approach to reporting varies across surveys. While some of the surveys made inferences at the country level and comparisons across countries, others refrained from doing so. Differences in the scope of inferences seem to be explained, at least partly, by the attributes of the sample (e.g. sampling strategy, size and skewness in terms of country representativeness) and the presence and form of post hoc adjustment made when processing the data.

The surveys based on quota samples (i.e. OECD Risks that Matter Survey 2020 and Trustlab survey) that had nationally representative samples according to certain demographic and socio-economic characteristics with at least 1 000 responses per country made inferences at the country level.

Although not relying on quota samples, the Eurofound Living, Working and COVID-19 e-survey, the Global Drug Survey 2021 and the EU-LGBTI II Survey, all with relatively large and balanced sample sizes across countries, also made inferences at the country level with certain conditions. The Eurofound Living, Working and COVID-19 e-survey reports country-level estimates by ensuring that each of these estimates are based on a minimum sample size (Ahrendt et al., 2020[33]; Sandor and D., 2020[34]). These minimum thresholds are based on an effective sample size of 100 for each question but also correcting for the demographic composition of the survey. The Eurofound Living, Working and COVID-19 e-survey also highlights estimates as having low reliability if the effective sample size is between 100 and 200. The EU-LGBTI II Survey, which managed to surpass its target sample size based on the LGBTI category and age group in almost all countries, also drew inferences at the country level (European Union Agency for Fundamental Rights, 2020[31]).

The T4 Teachers and Technology Global Survey, which applied a combination of convenience, purposive and snowballing sampling, mainly limited itself to reporting about the population of respondents. In the report, there are four countries (i.e. India, Nigeria, the Philippines and the United Arab Emirates) with more than 200 respondents for which country-level inferences are drawn (Pota et al., 2021[22]).

The surveys with relatively low sample sizes (i.e. below 3 000 respondents in total across 60-100 countries depending on the survey) and/or highly skewed samples in terms of country representativeness, tend to present results only in terms of the population respondents and avoid making inferences at the country level (i.e. Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19; Cedefop: COVID-19 pandemic and career guidance systems and policy development – Joint survey by international organisations, OECD Science Flash Survey 2020) (Table A.2). Although the UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education managed to reach a considerably larger sample size, as it was highly skewed towards certain countries, it also opted to report only about the population of respondents without providing country-level results.

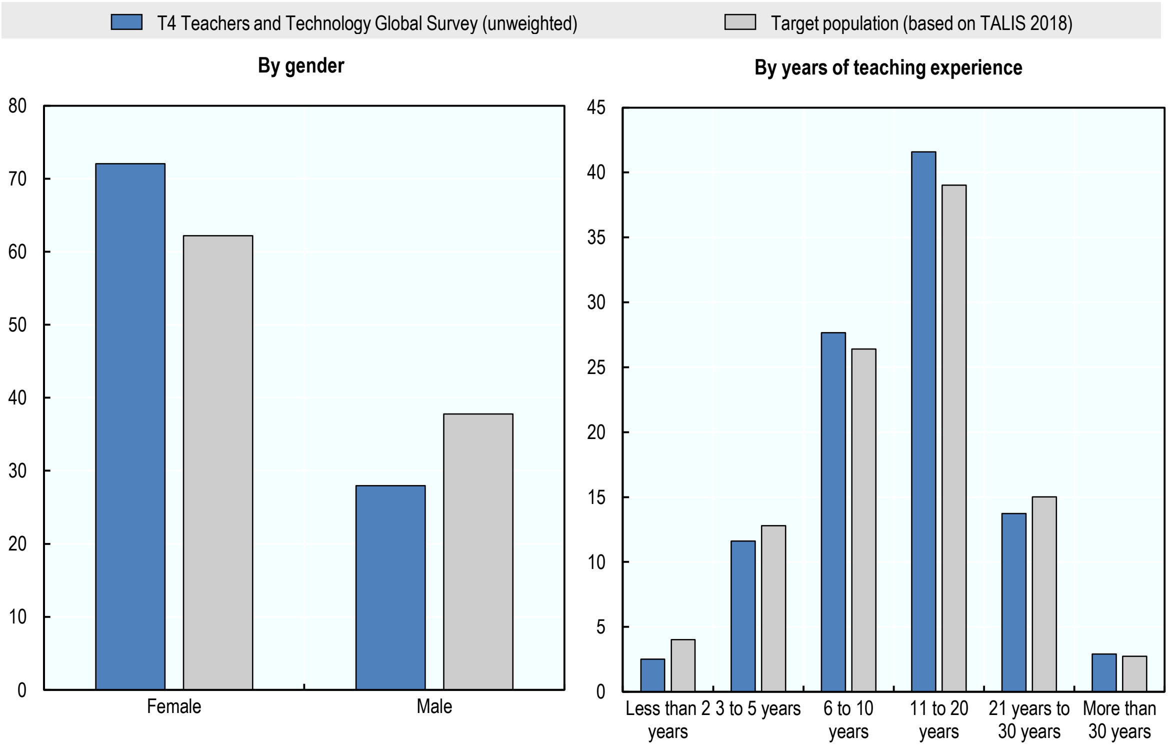
In the case of the surveys based on quota samples, the sample distributions match the target population with respect to certain demographic variables (e.g. age, gender, income) by design (i.e. OECD Risks that Matter Survey 2020, Trustlab survey).

As for the rest of the surveys that relied on convenience samples or a combination of convenience, purposive and snowballing samples, reporting such distributions is essential to providing important information on their non-probability samples. Nevertheless, reporting how the sample distributions match the target population along more easily accessible demographic variables or other theoretically grounded confounders is not a common practice across the surveys reviewed. There can be several reasons behind the lack of reporting about sample distributions in comparison to the target population. In some cases, the target population itself may not be well defined, covering a wide range of stakeholders related to the theme of the survey (e.g. Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19; Cedefop: COVID-19 pandemic and career guidance systems and policy development – Joint survey by international organisations, OECD Science Flash Survey 2020). In some other cases, the target population is hard to reach and as a result it may be challenging to assess its distribution even according to basic demographic variables (e.g. Global Drug Survey 2021, EU-LGBTI II Survey).

Nevertheless, whenever feasible, reporting if the sample distributions match the target population with respect to certain demographic variables can provide valuable information about the extent of potential bias of the analysis based on a non-probability sample. For instance, comparing the sample of teachers in the United Arab Emirates from the T4 Teachers and Technology Global Survey to the distribution of the target population with respect to gender and teaching experience, can reveal some potential sources of bias (Figure 3). The comparison of the distributions show that both female teachers and those with 6 to 20 years of teaching experience were somewhat over-represented in the T4 Teachers and Technology Global Survey compared to the target population. It is plausible that both gender and years of teaching experience are also correlated with the use of technology for teaching, which in turn could lead to bias.

Figure 3. Reporting on the extent to which sample distributions match the target population

Percentage of respondents and the target population in the United Arab Emirates, by age and years of teaching experience



Note: The United Arab Emirates was selected to illustrate the value of reporting about sample distributions since it was the only country with a large enough sample from the T4 Teachers and Technology Global Survey that was also covered by TALIS 2018. Data refer to lower secondary teachers (i.e. in the case of the T4 Teachers and Technology Global Survey sample lower secondary teachers are defined as those teaching 12-16-year-old students). The T4 Teachers and Technology Global Survey sampled 212 teachers teaching 12-16-year-old students across all the countries and territories.
Source: OECD (2018[35]), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 23 March 2022); and adapted from Pota, V. et al. (2021[22]), *T4 Turning to Technology: A Global Survey of Teachers' Responses to the Covid-19 Pandemic*, https://t4.education/t4-insights/turning-to-technology?hsLang=en (accessed on 23 March 2022).

Design-based inferential tools, such as confidence intervals, which are commonly used to make inferences about the general population, assume probabilistic sampling (i.e. a random process of data generation that can hypothetically be replicated). In the case of non-probability samples, the underlying probabilistic assumptions that would justify the use of these common tools are absent. Most of the reviewed surveys do not report margins of uncertainty in the absence of design-based inference methods. However, some of the non-probabilistic survey projects do report such margins (e.g. calculating statistical significance) based on an assumption of simple random sampling from an infinite population (e.g. Eurofound Living, Working and COVID-19 e-survey).

3.3. Review of transparency and replicability of data collection and data processing

Although transparency and the replicability of the survey projects reviewed were reported to be important objectives for most survey projects, the availability of methodological reports and documentation is limited in most cases. Often, the publicly available methodological information refer briefly to the non-probabilistic nature of the survey and hence the non-representativeness of the results, but do not go into detail about data collection and data processing, not to mention reporting more in detail about the characteristics of the non-probability sample.

Yet there are also some examples for more in-depth reporting about data collection and data processing. For instance, the EU-LGBTI II Survey is accompanied by a fully-fledged technical report that covers the whole survey process, from study design through data collection to data processing (European Union Agency for Fundamental Rights, 2020[31]). This type of extensive technical documentation is essential to providing transparency about the assumptions and methods used to recruit the sample, collect the data and make inferences.

3.4. Lessons learnt for TALIS from current practices in the use of non-probability samples in other comparative, cross-national surveys

3.4.1. Rationales for using non-probability samples in the context of TALIS

TALIS has a well-established, high-quality survey based on probability samples; therefore, the common rationales for using non-probability samples across the surveys reviewed, such as timeliness, feasibility and cost efficiency, need to be considered with the risks inherent in non-probabilistic approaches in mind. In the case of TALIS, compiling sampling frames for teachers and school leaders is feasible and the response rates tend to be high – for instance, at the lower secondary level, the overall response rate of teachers is around 85% on average across TALIS participants (OECD, 2020[13]). Relying on non-probabilistic sampling approaches in the context of TALIS would likely lead to a substantial loss in accuracy and the impossibility of quantifying uncertainty in findings.

Nevertheless, as some of the examples of the current practices of non-probability samples highlight, there can be scenarios where using a non-probability sample in the TALIS context can be justified and risks associated with the non-probability sampling can be mitigated.

There are certain types of questions and analyses for which non-probability samples are more fit than others. As shown by past research, non-probability samples are more suitable for analyses focusing on a narrow set of estimates, as these kinds of analyses require the control of a smaller set of covariates (Baker et al., 2013[1]). In addition, non-probability samples also fit better for exploring relationships among a broad set of characteristics rather than precisely describing those characteristics in the population of interest (Baker et al., 2013[1]). This is a point also made by the Global Drug Survey when advocating that unrepresentativeness is less of a concern if the interest of the survey lies in identifying trends and relationships rather than providing accurate point estimates about the phenomenon of interest (Barratt et al., 2017[32]). Nevertheless, it is important to highlight that the validity of an analysis exploring relationships among variables based on non-probability samples may require some empirical justification and it may not extend to all surveys and topics.

As shown by the Eurofound Living, Working and COVID-19 e-survey, non-probability samples can be a useful, cost-efficient vehicle for implementing innovative elements, such as a panel component that allows tracking of the evolution of the same respondents over time. Collecting panel data would have great potential for TALIS as it could allow for venturing into new types of analyses that are currently hindered by the cross-sectional nature of the TALIS survey (e.g. examining the same teachers to detect any changes in their practices, beliefs and well-being that might occur over a period of time).

A complementary non-probabilistic survey implemented within the TALIS umbrella could leverage the high-quality, probabilistic TALIS survey. In the review, the only survey that could build on well-established surveys based on probability samples was the Eurofound Living, Working and COVID-19 e-survey. As highlighted by this survey project, having such a link can bring various benefits from relying on high-quality questionnaire design and translation procedures to having access to a wide range of variables on the target population collected through probability sampling.

In addition, as the review of current uses of non-probability samples in cross-national surveys shows, non-probability samples can also serve communication goals and be used as engagement and outreach tools. Reporting on new information generated from non-probability samples can help in raising awareness of certain issues and topics as well as in promoting other related work. Yet it is important to clearly distinguish communication goals from genuine knowledge generation and to emphasise the limitations and risks associated with the use of non-probability samples.

3.4.2. Risks associated with the use of non-probability samples and potential mitigation strategies in the context of TALIS

Similar to all survey projects included in the review, the main sources of risks of using non-probability sampling in the context of TALIS are non-coverage and uncontrolled selection that can lead to biased estimates. For instance, if TALIS were to implement an open-link web survey of teachers, or build an online access panel through quota sampling (by invitation), it would inevitably exclude teachers who have either limited access to the Internet or have a low online presence. Excluding these teachers from a survey would lead to biased estimates if these teachers were to differ from those who have adequate access to the Internet and have considerable online presence on characteristics that are likely to be of interest in the survey (i.e. bias due non-coverage). Selection bias would occur if there were differences between teachers who were supposed to be represented in the survey (e.g. teachers who could be reached on line) and teachers who not only had access to the survey but also decided to take part in it. One can never be sure about the accuracy of inferences to the general population of teachers, in particular if these generalisations concern point estimates that are based on a non-probability sample.

As the literature and the present review of survey projects relying on non-probability samples show, there are strategies available to mitigate the risks inherent in non-probability sampling. These strategies are related to various stages of the data cycle: study design, data collection, data processing and reporting.

It is important to think of variables that are most likely correlated with both inclusion in the sample and the outcome of interest (i.e. theoretically grounded confounding variables) already at the study design phase, so that information on these variables can be collected at a later stage (Mercer et al., 2017[10]). Having information on confounding variables allows for the reduction of risks, both at the data processing (i.e. weighting) and reporting (i.e. distributional information of the sample and target population with respect to the confounding variables) stages. Although collecting information on all potential confounders is not possible in practice, the more are accounted for at the survey design phase and, hence collected later on, the better the chances of reducing risks associated with non-probability sampling.

If TALIS were to administer an online open-link survey of teachers, it would have to already consider at the survey design phase how to limit the risk of people outside the target population participating in the survey. Failing to do so would lead to potential bias. Various survey projects reviewed in earlier sections addressed this issue through the design of the questionnaire, including conditional screener questions. Although questionnaire design alone cannot fully exclude the potential bias that can be caused by the participation of people outside the target population, because respondents can lie to screener questions, it does reduce such risks.

Besides focusing on the inherent risks of non-probability samples at the study design phase, it is equally important to ensure that the broader survey methodology, including the questionnaire design and the related translation processes, is of high quality. If a non-probabilistic TALIS survey was implemented, relying on the existing TALIS questionnaires and their translations would ensure the validity of the constructs and reduce the risks of measurement error.

Based on the experience of the online open-link surveys reviewed, real-time monitoring of the sample size and composition and fine-tuning of the recruitment strategies (e.g. social media advertisements) during data collection is desirable. Such an approach to data collection can reduce the risks associated with bias due to non-coverage and increase the sample size.

After study design and data collection, one can still attempt to mitigate the two main risks inherent in non-probability sampling – bias due to non-coverage and bias due to uncontrolled selection – when processing the data. Post hoc adjustments typically refer to applying weights to the sample of respondents so that they resemble the target population more closely. Although various surveys included in this review applied such adjustments to reduce the potential bias inherent in non-probability samples, it is important to note that these techniques are unlikely to eliminate all potential biases, they can only reduce it at best (Tourangeau, Conrad and Couper, 2013[7]).

Similar to the open-link web surveys reviewed, the chances of certain participants responding more than once to manipulate the outcome of a non-probabilistic TALIS survey can be considered negligible given the time required to complete the survey and the likely impact of such non-genuine responses. Yet, as TALIS results can inform policy decisions that may affect the respondents (i.e. teachers and school leaders) themselves, the risk of non-genuine responses cannot be fully excluded either. To address this issue at the data processing stage, one option is to try to ensure single entry by each respondent through removing duplicate observations using multiple conditions.

Reporting is the last phase in the survey cycle when risks associated with non-probability samples can be mitigated. At this stage, risk reduction refers to acknowledging the limitations of non-probability sampling by providing as much information as possible about the characteristics and potential bias of the non-probability sample collected. Similar to some of the survey projects reviewed, the scope of inferences (e.g. overall population of respondents vs. country level) drawn from a non-probability sample should be conditional on sample size requirements,[11] post hoc adjustments and distributional information about the sample and target population with respect to the confounding variables.

Being clear about the limitations of non-probability sampling also requires that the reporting of results and subsequent inferences be faithful to the methodological limitations inherent in non-probability samples. Thus, similar to the approach of most of the survey

[11] The Eurofound Living, Working and COVID-19 e-survey draws country-level inferences if the effective sample size, which corrects for the composition of the survey by taking into account the design effect due to weighting (based on the Kish formula), reaches a certain threshold (i.e. 100 and 200). Adjusting for the composition of the survey is particularly important in the case of an online open-link survey, since responses can be skewed towards particular demographic groups or countries.

projects reviewed, an eventual non-probabilistic survey conducted within TALIS should not report margins of uncertainty in the absence of design-based inference methods.

Although not a common practice across the reviewed surveys, examining and reporting if the sample distributions match the target population with respect to more simply accessible demographic variables (gender, age, etc.), as well as other confounders, provides valuable insights about the characteristics and potential bias of a non-probability sample. This is all the more important given the fact that a non-probability sample within TALIS would have access to a wide range of information about teachers and schools leaders as a result of the core TALIS survey.

4. Conclusions: potential roles non-probability samples could play in the context of TALIS

4.1. Complementary module with a longitudinal design exploring trends and relationships

Non-probability sampling could contribute to TALIS through a complementary module with a longitudinal design. A panel of volunteer teachers or school leaders, among the sampled participants of the core TALIS survey, who agree to be contacted again in the future, could participate in follow-up surveys by providing their contact details. Collecting panel data would allow the examination of the same teachers or school leaders to detect any changes in their practices, beliefs and well-being that might occur over a period of time. Exploring relationships (between baseline and follow-up levels of a same variable, and across constructs) rather than aiming to provide accurate point estimates within such a non-probabilistic complementary module is preferred, as the unrepresentative nature of the sample may be less problematic with this type of analysis.

For instance, in conjunction with the Global Teaching InSights: A Video Study of Teaching, which is an international large-scale, video-based study of teaching and learning with a probabilistic sampling design (OECD, 2020[36]), a complementary, non-probabilistic TALIS module with a longitudinal design could also explore how certain teaching practices are interrelated, and how contextual aspects of teaching are related to teacher characteristics.

Such a complementary longitudinal module could focus, for instance, on analysing the effect of support mechanisms on the practices, beliefs and well-being of novice teachers in the first few years of their career. In order to ensure that a sufficient number of novice teachers are recruited, a referral mechanism might need to be considered under this scenario.[12]

A complementary TALIS module based on a non-probability sample should rely, to the extent possible, on the core TALIS survey. This would allow the high-quality TALIS instruments and translation procedures to be harnessed. In addition, it could also help by leveraging the wide range of variables collected by the core TALIS survey in order to reduce the risks associated with non-probability sampling and to assess bias in the non-probabilistic component.

[12] For example, if TALIS respondents provide their own contact details, they could also be sent a code or unique participation link that they could share (privately) with a number of novice colleagues. This would require adequate protection for respondents' privacy, credible screening procedures to ensure that newly recruited respondents are in scope, and potentially incentives for the original respondents who provide quality referrals.

4.2. Field trial data collection

The field trial data collection, which is implemented to validate the TALIS instruments and derived measures and to trial the operational procedures, follows a probabilistic design. That is, the field-trial data collected in each participant country are representative of each education system. However, non-probability samples could play a role in the development phase (e.g. assessing model fit and item parameters for item-response-theory models or piloting survey instruments) of the core TALIS survey that itself is based on probabilistic sampling. Relying on non-probabilistic sampling for the field trial could lead to more timely access to field-trial data. Yet it is important to ensure that the operational component of the field trial, whose main goal is to rehearse the sampling and survey operations ahead of the main study, is fulfilled.

Using probability samples for piloting and field trial data collection is not necessarily the case for other international large-scale surveys that have a probabilistic design. For instance, the field trial data collection of the Programme for International Student Assessment (PISA) relies on non-probability samples while fulfilling both the operational and the instrument-development components of the field trial. The PISA field trial assures that the school sample assembled by the national centres (NCs) is sufficiently diverse, does not exclude significant parts of the target population, and has sufficient students to validate the instruments, etc. Even though the sampling procedures of the field trial and main survey data collections are different, the PISA sampling contractor ensures that NCs can rehearse during the field trial for the main survey sampling procedure. This is done by administering and verifying the same list of sampling tasks during the field trial that are present in the main survey data collection.

4.3. Exploratory survey with scoping and outreach purposes

Moving away from generalisability, which tends to be a central goal of survey research, non-probabilistic sampling approaches can be used to learn about the existence of some unknown phenomenon or to gain some very preliminary understanding of a phenomenon. More concretely, in the case TALIS, such an exploratory survey based on non-probability sampling could be used to identify new topics that are deemed relevant by the teaching profession to include in the survey. Such an exploratory survey could take an open-ended format, reflecting on the nature of the survey, and could result in the development of new questionnaire items for the main survey or new thematic modules. In addition, this type of exploratory survey would also provide a vehicle for a more direct and timely dialogue between the wider teaching profession and policy makers. It could function as an engagement and outreach tool by allowing teachers to take a more active role in shaping TALIS and, as a result, school and teacher policies.

Annex A. Background information about the review and the surveys included in the review

Table A.1. Questions used for the review of the non-probabilistic surveys covered

	Questions
Review of rationale for using non-probability samples	What was the rationale for using non-probability samples instead of a probability sample?
	Was there a similar survey based on a probability sample that you built on, or were you intending to administer the survey to a probability sample later on? What is the relationship between the two surveys?
	In hindsight, what do you think were the main advantages and challenges of using a non-probability sample, in particular, in terms of accuracy, credibility (impact and reputation), timeliness, cost efficiency and replicability of the study?
Review of risks associated with the use of non-probability samples and associated mitigation strategies at various stages of the data cycle	During study design, did you anticipate the extent to which your non-probability samples could be subject to potential bias, in particular due to non-coverage and uncontrolled selection? If yes, …
	In the case of your survey, what could be the source(s) of non-coverage? Did you consider alternative recruitment strategies for your sample or administration modes and their possible effect on coverage?
	In the case of your survey, did you think of variables that are most likely correlated with both inclusion in the sample and the outcome of interest (i.e. theoretically grounded confounding variables)? If yes, did you make efforts to collect information on them?
	In the case of opt-in/open-link/volunteer surveys, did you consider the following risks? How did you address these (in the design of questionnaires and screener questions, when promoting the survey to potential participants, in the processing of data, etc.)?
	the risk of people outside the target population participating in the survey
	the risk of participants responding more than once (on the same, or on a different computer)
	the risk of a group of people manipulating the outcomes of the survey
	During data collection, did you use a quota strategy, or otherwise monitor sample size and composition and use targeted recruitment strategies (advertisements, incentives) to address any emerging issues?
	Did you make an attempt during data processing to mitigate the risks associated with the use of non-probability samples?

	Questions
	Did you use any method after data collection that involved adjusting the weights assigned for the survey participants to remove or reduce bias (e.g. post-stratification, raking, generalised regression [GREG] modelling, propensity score matching)?
	When reporting results, …
	How did you acknowledge the limitations of non-probability sampling?
	Did you report sample distributions (before and after weighting) of theoretically grounded confounders against their distribution in the target population?
	Did you examine and report if the sample distributions match the target population with respect to more simply accessible demographic variables (gender, age, etc.)?
	How did you compute and report margins of uncertainty in the absence of design-based inference methods?
Review of transparency and replicability of data collection and data processing	Was the replicability of the study an important objective? In particular, to what extent do reports and technical documentation provide transparency about the assumptions and methods used to recruit the sample, collect the data and make inferences?
Are there any other cross-national surveys relying on non-probabilistic sampling methods that you are aware of? If yes, could you share their names?	

Table A.2. Overview of the non-probabilistic surveys included in the review

	Context	Data collection	Number of responses[1]	Number of countries	Sampling method	Eligibility criteria defining study population	Reporting goals and the scope of inferences
OECD Risks that Matter Survey 2020	Societies' social and economic concerns as well as expectations about social protection systems	September-October 2020	Over 25 000 (minimum sample of 1 000 respondents per country)	25 OECD countries	Quota sample. Nationally representative by gender, age group, education level*, income level, and worker status. Implemented on line by a private polling company.	People aged between 18 and 64	Inferences about the target population by country. Comparisons across countries.

	Context	Data collection	Number of responses[1]	Number of countries	Sampling method	Eligibility criteria defining study population	Reporting goals and the scope of inferences
Trustlab survey	Social preferences, generalised trust and trust in institutions using experimental games	Between Nov 2016 and Nov 2017	6 320 (minimum. sample of 1 000 respondents per country)	6 OECD countries	Quota sample. Nationally representative by gender, age group and income level. Implemented by private polling company.	People aged 18 or above	Inferences about the target population by country. Comparisons across countries.
Joint OECD-Harvard Graduate School of Education survey on the effects of COVID-19	Information on the education conditions faced in countries, and on the approaches adopted to sustain educational opportunity amidst the COVID-19 pandemic	April-May 2020	1 370 respondents (teachers among others)	59 countries	Combination of convenience, purposive and snowballing sample. Online open-link questionnaire.	Teachers, school principals, senior government officials, education administrators, employees in education companies, etc.	Reporting only about the population of respondents. No disaggregation by country.
CEDEFOP: COVID-19 pandemic and career guidance systems and policy development – Joint survey by international organisations	Snapshot of career guidance delivery, services, usage and careers learning in countries during the COVID-19 pandemic	June-August 2020	963 respondents	93 countries	Combination of convenience, purposive and snowballing sample.	People working for national, regional and other bodies responsible for managing career guidance services, guidance practitioners and researchers, etc.	Reporting only about the population of respondents. No disaggregation by country.
OECD Science Flash Survey 2020	Barometer of the state of science amidst the COVID-19 crisis	Between April 2020 and June 2021	Almost 3 000 respondents	Around 100 countries	Convenience sample. Online open-link questionnaire.	Scientists or any other individuals with an interest in science or science policy (only available in English)	Reporting only about the population of respondents. No disaggregation by country.

	Context	Data collection	Number of responses[1]	Number of countries	Sampling method	Eligibility criteria defining study population	Reporting goals and the scope of inferences
OECD International Survey of Science (ISSA) 2021	Key topical issues for the global scientific community, focusing on scientists' work conditions, engagement with society and the impact of the COVID-19 on their work and career prospects	Between April 2021 and January 2022	Over 3 000 respondents	81 countries	Convenience sample. Online open-link questionnaire with ORCID authentication (75% of the sample) or email authentication (25% of the sample).	Researchers, regardless of the type of institution	Inferences about the target population by country. Main focus on relationships between structural, demographic and behavioural variables.
T4 Teachers and Technology Global Survey	Impact of the pandemic on teachers and learners globally, focusing on technology use and digital resources	April-May 2021	20 679 respondents (i.e. teachers)	165 countries	Convenience sample. Online open-link questionnaire.	Teachers	Reporting mainly about the population of respondents. Country-level inferences only for four countries (with more than 200 respondents).
UNESCO Global Survey for Teachers on Education for Sustainable Development and Global Citizenship Education	Teachers' readiness to teach themes around "learning to live sustainably" and "learning to live together in peace" - in relation to SDG Target 4.7, which calls for countries to "ensure that all learners acquire the knowledge and skills needed to promote sustainable development"	March-April 2021	58 280 respondents (i.e. teachers)	144 countries	Convenience sample. Online open-link questionnaire.	Teachers	Reporting only about the population of respondents. No disaggregation by country.

	Context	Data collection	Number of responses[1]	Number of countries	Sampling method	Eligibility criteria defining study population	Reporting goals and the scope of inferences
Eurofound Living, Working and COVID-19 e-survey	Economic and social effects (i.e. impact on well-being, health and safety, work and telework, people's work-life balance and financial situation) of the COVID-19 pandemic in the European Union	Three rounds: April 2020, July 2020 and March 2021	138 629 respondents (across the three rounds: R1 = 67 685; R2 = 24 143; R3 = 46 800)	27 member states of the European Union	Convenience sample. Online open-link questionnaire. Sample re-weighted and is nationally representative by gender, age, education level and self-defined urbanisation levels.	People aged 18 or above	Inferences about the target population by country. Comparisons across countries.
Global Drug Survey 2021	Early identification of new drug trends and drug-related harm	November-December 2020	32 022 respondents	22 countries	Combination of convenience, purposive and snowballing sample. Online open-link questionnaire.	People aged above 16	Inferences about the target population by country. Comparisons across countries and over time.
European Union Agency for Fundamental Rights' 2019 LGBTI survey (EU-LGBTI II Survey)	Experience of LGBTI people with discrimination, violence and harassment in various areas of life, including employment, education, healthcare, housing and other services	May-July 2019	139 799 respondents	27 member states of the European Union, the United Kingdom, North Macedonia and Serbia	Combination of convenience, purposive and snowballing sample. Online open-link questionnaire.	People aged 15 or above who describe themselves as lesbian, gay, bisexual, trans or intersex (LGBTI)	Inferences about the target population by country. Comparisons across countries and over time.

Notes: 1. Number of responses after data cleaning.

References

Ahrendt, D. et al. (2020), "Living, working and COVID-19: Methodological Annex to Round 1", *Working Paper*, No. WPEF20005, Eurofound, Dublin, https://www.eurofound.europa.eu/sites/default/files/wpef20005.pdf (accessed on 24 March 2022). [33]

Ansolabehere, S. and B. Schaffner (2014), "Does survey mode still matter? Findings from a 2010 multi-mode comparison", *Political Analysis*, Vol. 22/3, pp. 285-303, https://doi.org/10.1093/pan/mpt025. [18]

Baker, R. et al. (2013), "Summary report of the AAPOR Task Force on Non-probability Sampling", *Journal of Survey Statistics and Methodology*, Vol. 1/2, pp. 90-143, https://doi.org/10.1093/jssam/smt008. [1]

Barratt, M. et al. (2017), "Moving on from representativeness: Testing the utility of the Global Drug Survey", *Substance Abuse: Research and Treatment*, Vol. 11, pp. 1-17, https://doi.org/10.1177/1178221817716391. [32]

Beaumont, J. (2020), "Are probability surveys bound to disappear for the production of official statistics?", *Survey Methodology*, Vol. 46/1, pp. 1-28, https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X202000100001. [12]

Bethlehem, J. (2017), "Nonprobability online panels", in *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, Chichester, UK, https://doi.org/10.1002/9781118445112.stat08076. [20]

Bethlehem, J. (2016), "Solving the nonresponse problem with sample matching?", *Social Science Computer*, Vol. 34/1, pp. 59-77, https://doi.org/10.1177/0894439315573926. [14]

Brüggen, E., J. van den Brakel and J. Krosnick (2016), "Establishing the accuracy of online panels for survey research", *Discussion Paper*, No. 04, Statistics Netherlands, The Hague, https://www.cbs.nl/en-gb/background/2016/15/establishing-the-accuracy-of-online-panels-for-survey-research (accessed on 21 April 2022). [16]

Cedefop; European Commission; ETF; ICCDPP; ILO; OECD; UNESCO (2020), *Career guidance policy and practice in the pandemic: Results of a joint international survey June to August 2020*, Publications Office of the European Union, Luxembourg, http://data.europa.eu/doi/10.2801/318103 (accessed on 24 March 2022). [26]

Cornesse, C. et al. (2020), "A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research", *Journal of Survey Statistics and Methodology*, Vol. 8/1, pp. 4-36, https://doi.org/10.1093/jssam/smz041. [6]

Dutwin, D. and T. Buskirk (2017), "Apples to oranges or Gala versus Golden Delicious? Comparing data quality of nonprobability Internet samples to low response rate probability samples", *Public Opinion Quarterly*, Vol. 81, pp. 213-249, https://doi.org/10.1093/poq/nfw061. [15]

Eurofound (2021), "Living, working and COVID-19 (Update April 2021): Mental health and trust decline across EU as pandemic enters another year", *Factsheet*, No. EF/21/064/EN, Publications Office of the European Union, Luxembourg, https://www.eurofound.europa.eu/publications/report/2021/living-working-and-covid-19-update-april-2021-mental-health-and-trust-decline-across-eu-as-pandemic (accessed on 23 March 2022). [27]

Eurofound (2020), *Living, Working and COVID-19*, COVID-19 series, Publications Office of the European Union, Luxembourg, https://doi.org/10.2806/76040. [28]

European Union Agency for Fundamental Rights (2020), *A Long Way to Go for LGBTI Equality - EU-LGBTI II*, Publications Office of the European Union, Luxembourg, https://doi.org/10.2811/582502. [29]

European Union Agency for Fundamental Rights (2020), *A Long Way to Go for LGBTI Equality - EU-LGBTI II: Technical Report*, Publications Office of the European Union, Luxembourg, https://doi.org/10.2811/447153. [31]

Foote, C. et al. (2021), "Measuring the U.S. Employment Situation Using Online Panels: The Yale Labor Survey", *Discussion Paper*, No. 2282, https://cowles.yale.edu/sites/default/files/files/pub/d22/d2282.pdf (accessed on 8 June 2021). [5]

Goel, S., A. Obeng and D. Rothschild (2015), "Non-representative surveys: Fast, cheap, and mostly accurate", *Working Paper*, No. 27, http://researchdmr.com/FastCheapAccurate.pdf. [2]

Graefe, A. (2016), "Forecasting proportional representation elections from non-representative expectation surveys", *Electoral Studies*, Vol. 42, pp. 222-228, https://doi.org/10.1016/j.electstud.2016.03.001. [3]

Groves, R. et al. (2009), *Survey Methodology, Second Edition*, Wiley Series in Survey Methodology, John Wiley & Sons, Inc., Hoboken, NJ. [9]

MacInnis, B. et al. (2018), "The accuracy of measurements with probability and nonprobability survey samples: Replication and extension", *Public Opinion Quarterly*, Vol. 82/4, pp. 707-744, https://doi.org/10.1093/poq/nfy038. [17]

Mercer, A. et al. (2017), "Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference", *Public Opinion Quarterly*, Vol. 81/S1, pp. 250-271, https://doi.org/10.1093/poq/nfw060. [10]

Murtin, F. et al. (2018), "Trust and its determinants: Evidence from the Trustlab experiment", *OECD Statistics Working Papers*, No. 2018/2, OECD Publishing, Paris, https://dx.doi.org/10.1787/869ef2ec-en. [25]

OECD (2021), *Main Findings from the 2020 Risks that Matter Survey*, OECD Publishing, Paris, https://dx.doi.org/10.1787/b9e85cf5-en. [24]

OECD (2020), *Global Teaching InSights: A Video Study of Teaching*, OECD Publishing, Paris, https://dx.doi.org/10.1787/20d6f36b-en. [36]

OECD (2020), "Schooling disrupted, schooling rethought: How the Covid-19 pandemic is changing education", *OECD Policy Responses to Coronavirus (COVID-19)*, OECD Publishing, Paris, https://dx.doi.org/10.1787/68b11faf-en. [21]

OECD (2020), "Technical notes on sampling procedures, response rates and adjudication for TALIS 2018", in *TALIS 2018 Results (Volume II): Teachers and School Leaders as Valued Professionals*, OECD Publishing, Paris, https://doi.org/10.1787/baeccd55-en. [13]

OECD (2018), *TALIS 2018 Database*, https://www.oecd.org/education/talis/talis-2018-data.htm (accessed on 23 March 2022). [35]

OECD (2011), *Quality Framework for OECD Statistical Activities*, https://www.oecd.org/sdd/qualityframeworkforoecdstatisticalactivities.htm (accessed on 24 July 2021). [8]

OECD (2008), *OECD Glossary of Statistical Terms*, OECD Publishing, Paris, https://doi.org/10.1787/9789264055087-en. [38]

Pasek, J. (2015), "When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence", *International Journal of Public Opinion Research*, Vol. 28/2, pp. 269-291, https://doi.org/10.1093/ijpor/edv016. [19]

Pota, V. et al. (2021), *T4 Turning to Technology: A Global Survey of Teachers' Responses to the Covid-19 Pandemic*, T4 Education, Hemel Hempstead, https://t4.education/t4-insights/turning-to-technology?hsLang=en (accessed on 23 March 2022). [22]

Sandor, E. and A. D. (2020), "Living, working and COVID-19: Methodological Annex to Round 2", *Working Paper*, No. WPEF20023, Eurofound, Dublin, https://www.eurofound.europa.eu/sites/default/files/wpef20023.pdf (accessed on 24 March 2022). [34]

Schaurer, I. and B. Weiß (2020), "Investigating selection bias of online surveys on coronavirus-related behavioral outcomes", *Survey Research Methods: Journal of the European Survey Research Association*, Vol. 14/2, pp. 103-108, https://doi.org/10.18148/SRM/2020.V14I2.7751. [11]

Tourangeau, R., F. Conrad and M. Couper (2013), *The Science of Web Surveys*, Oxford University Press, Oxford, https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199747047.001.0001/acprof-9780199747047. [7]

UNESCO; Education International (2021), *Teachers Have Their Say: Motivation, Skills and Opportunities to Teach Education for Sustainable Development and Global Citizenship*, United Nations Educational, Scientific and Cultural Organization, Paris and Education International, Brussels, https://unesdoc.unesco.org/ark:/48223/pf0000379914 (accessed on 23 March 2022). [23]

Wang, W. et al. (2015), “Forecasting elections with non-representative polls”, *International Journal of Forecasting*, Vol. 31/3, pp. 980-991, https://doi.org/10.1016/j.ijforecast.2014.06.001. [4]

Winstock, A. and J. Ferris (2020), “Nitrous oxide causes peripheral neuropathy in a dose dependent manner among recreational users”, *Journal of Psychopharmacology*, Vol. 34/2, pp. 229-236, https://doi.org/10.1177/0269881119882532. [37]

Winstock, A. et al. (2021), *Global Drug Survey (GFDS) 2021 Key Findings Report*, https://www.globaldrugsurvey.com/wp-content/uploads/2021/12/Report2021_global.pdf (accessed on 21 March 2022). [30]