ANNUAL REVIEWS

Annual Review of Organizational Psychology and Organizational Behavior

Innovations in Sampling: Improving the Appropriateness and Quality of Samples in Organizational Research

Michael J. Zickar and Melissa G. Keith

Department of Psychology, Bowling Green State University, Bowling Green, Ohio, USA; email: mzickar@bgsu.edu

Annu. Rev. Organ. Psychol. Organ. Behav. 2023. 10:315–37

First published as a Review in Advance on October 31, 2022

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

https://doi.org/10.1146/annurev-orgpsych-120920-052946

Keywords

sampling, sampling techniques, online samples, data quality, insufficient effort responding, participant dishonesty

Abstract

Technology has changed the way that organizational researchers obtain participants for their research studies. Although technology has facilitated the collection of large quantities of data through online platforms, it has also highlighted potential data quality issues for many of our samples. In this article, we review different sampling techniques, including convenience, purposive, probability-based, and snowball sampling. We highlight strengths and weaknesses of each approach to help organizational researchers choose the most appropriate sampling techniques for their research questions. We identify best practices that researchers can use to improve the quality of their samples, including reviewing screening techniques to increase the quality of online sampling. Finally, as part of our review we examined the sampling procedures of all empirical research articles published in the *Journal of Applied Psychology* in the past 5 years, and we use observations from these results to make conclusions about the lack of methodological and sample diversity in organizational research, the overreliance on a few sampling techniques, the need to report key aspects of sampling, and concerns about participant quality.

INTRODUCTION

In recent years, the availability and relatively low cost of professional online samples have caused their use as a data collection tool to skyrocket. The rapid ascendance of online professional samples has inspired researchers to ask significant questions about sample appropriateness and quality (e.g., Aguinis et al. 2021, Bernerth et al. 2021). This type of scrutiny is productive and has resulted in screening techniques and strategies to improve the quality of all samples. However, such reliance, and concern about overreliance on one strategy, is nothing new. In previous decades, researchers were concerned about overreliance on college sophomores from university research pools (e.g., Gordon et al. 1986, Sears 1986). In reading some of those criticisms, we note that researchers raised similar types of questions and concerns about relevance and generalizability then as they do now.

In addition to concerns about online samples, Bergman & Jean (2016) lament that industrial–organizational psychology samples tend to underrepresent wage earners, low- and medium-skilled workers, and contract workers while overrepresenting managers, salaried employees, and executives. They argue that our overrepresentation of certain types of participants harms our science by not adequately representing the full range of the workplace. Because of this biased representation, they argue, we are unable to understand the boundary conditions of our results and theories, and we overlook research questions that may be relevant to individuals who fall outside our typical samples. Their argument is that organizational researchers need to do a better job of sampling a wider range of individuals to help improve our science and practice. On the basis of these different critiques, it is clear that organizational researchers need to be more deliberative in evaluating whether their sampling techniques are appropriate for their research questions and reaching the population of interest (i.e., the individuals or groups the researcher intends to draw conclusions about). We hope that this article provides suggestions to organizational researchers to improve the appropriateness and quality of their research samples.

In this article, we review different types of sampling techniques, many of which are infrequently used in organizational behavior research. Next, we identify types of research questions that are most appropriately targeted for a particular sampling technique. We also distinguish between sampling strategies and data collection methods, two ideas that are often confounded. We then discuss ways researchers may improve the quality of the data collected and how these mechanisms may differ according to the sampling strategy and data collection method. In addition, we present a survey of recent issues of the *Journal of Applied Psychology* to investigate the use of sampling techniques so as to illustrate best practices and opportunities for organizational researchers to improve.

REVIEW OF DIFFERENT SAMPLING TECHNIQUES

To provide an overview of the prevalence of each sampling technique and sampling choices more generally, we manually reviewed every article published in the *Journal of Applied Psychology* between January 2017 and December 2021. For each article, we separately coded each sample for the following characteristics: (*a*) sampling technique (convenience sample, snowball sample, probability-based/stratified sample, purposive sample, other); (*b*) sample source [college students, MBA or graduate students, field (single organization), field (multiple organizations), archival data, Qualtrics panel, Amazon's Mechanical Turk (MTurk) or CloudResearch, Prolific, social media, other]; and (*c*) data collection method (online, paper and pencil, observation, interview, case study, other, unknown). Studies that were not empirical or did not include primary samples (e.g., meta-analyses, simulated data) were not included. The review resulted in 729 different samples. **Table 1** presents a breakdown of the frequency of each of the coded characteristics. Throughout this

Table 1 Sampling characteristics of *Journal of Applied Psychology* articles (2017–2021)

	Characteristic	*k*	Percentage
Sampling technique	Convenience sampling	654	90.7%
	Snowball sampling	32	4.4%
	Probability-based sampling (including stratified)	14	1.9%
	Purposive sampling	5	0.7%
	Other	16	2.2%
Sample participants	Mechanical Turk/CloudResearch	135	19.3%
	Field (single organization)	113	16.2%
	Undergraduate college students	109	15.6%
	Field (multiple organizations)	107	15.3%
	Archival data	62	8.9%
	Graduate students/MBA students	30	4.3%
	Qualtrics panel	23	3.3%
	Prolific	23	3.3%
	Social media	16	2.3%
	Other	80	11.5%
Data collection	Online	654	90.7%
	Paper survey	32	4.4%
	Case study	16	2.2%
	Observation	14	1.9%
	Interview	5	0.7%

article, we refer to these results to identify areas of concern and to highlight strengths in organizational research. First, we define each of the five approaches to sampling.

Probability-Based Sampling

With probability-based sampling (PBS), a population is enumerated and individuals within the population are chosen on the basis of a stochastic process that gives each individual an equal chance of being chosen. The primary advantage of this approach is that it is mathematically possible to determine the statistical confidence intervals related to the estimate of population values. PBS is used frequently in domains where it is important to accurately estimate population values and one cannot afford to assess everyone in a population (e.g., political polling, opinion surveys, demography studies within the field of sociology). The challenge with PBS is accurately enumerating all members within a population so that you can choose your sample. In political polling, it is possible to do so because all voters must be registered before voting and voter lists are in the public record. PBS is infrequently used in organizational research, with only 1.9% of the samples in our review using a version of PBS. This is likely because such population lists are often difficult to obtain (e.g., an organization may be reluctant to provide names of all employees) or simply do not exist (e.g., you are interested in generalizing to all restaurant servers in the USA).

Stratified random sampling is a specialized case of PBS in which specific subgroups within the general population are identified and, within those strata, individuals are sampled to make sure there is sufficient representation of each subgroup, with everyone in a stratum given an equal chance of being selected. In most cases, individuals within small strata are oversampled. For example, Asian American female employees in the military may each be chosen at a higher probability than white male employees, given the relatively few military employees in the former group versus the latter. This technique is particularly useful if there needs to be an accurate estimation of minority groups that might not have sufficient representation if traditional PBS is used.

Convenience Sampling

Convenience sampling is a general term that indicates that the participants in the sample were chosen on the basis of ease of access or availability. This particular approach is most frequently used in organizational research, with 90.7% of the coded samples using convenience samples. With the availability of Internet platforms that provide online participants for researchers' data collections in return for a fee (paid to the participants and the online platform), platforms such as MTurk, Prolific, and Qualtrics panels are now common ways to collect samples (25.9% of all studies we coded used one of these types of services). Other examples of convenience sampling include administering a survey to students in a classroom (assuming that your classroom is not the population for which you wish to generalize), getting participants from a university-based subject pool, or administering surveys to an organization for which a friend works. The key for the convenience sample is that participants are chosen because of their willingness to respond to the survey itself. The challenge with convenience sampling is that it is unclear to what population the sample can be compared. With screening questions, however, online convenience samples can be targeted to meet certain criteria (e.g., people who work and/or make hiring decisions, respondents who have experienced a traumatic event recently), though the problem of the population of generalization still exists.

Purposive Sampling

Purposive sampling is an approach that acknowledges that researchers use their own judgment and expertise to choose participants. Etikan et al. (2016) identified several variants of purposive sampling, several of which we highlight here:

1. Expert sampling, where participants may be chosen because they have a particular expertise or knowledge related to the topic of interest.
2. Maximum variance sampling, where researchers choose individuals to cover a spectrum of beliefs or experiences.
3. Extreme case sampling, in which researchers choose a sample of individuals who are unusual (i.e., outliers in a population), such as heart attack patients who recovered much more quickly than the normal population.

An advantage of purposive sampling is that the participants are chosen to be of maximal interest to the researchers; disadvantages are that (*a*) it is difficult to determine what populations to which these samples generalize and (*b*) purposive sampling requires accurate identification of individuals that fit the purpose of the study. This sampling approach was used rarely (0.7%) in our coded studies.

Snowball Sampling

Snowball sampling relies on individual participants to identify other participants who are likely to meet a particular sample's inclusion criteria and to pass along study information to these potential participants. Just like a snowball gathers mass as it rolls down a snowy hill (at least in the cartoons), individual participants help increase the sample size by passing the survey or its link on to other potential participants. This technique is particularly useful for studying individuals who belong to hard-to-identify groups. For example, for a study that looks at transgender employees, there is no formal association of transgender employees and an individual's identity may be cloaked to other employees, so snowball sampling may be the only way to get a sample large enough to compute statistics. Limitations of this approach are that it is unclear how samples relate to populations of interest and that snowball sampling relies on the goodwill of participants, so bias may be induced

in the method such that individuals likely to recruit others to participate may be different from individuals less likely to pass along a survey (Marcus et al. 2017). Snowball sampling was used infrequently (4.4%) in the studies that we coded.

GENERAL OBSERVATIONS ABOUT THE SAMPLING TECHNIQUES

Several key observations need to be made about the five sampling techniques presented here. First, PBS methods, which include stratified random sampling, can be distinguished from non-PBS approaches, which include convenience, purposive, and snowball sampling. In the latter three approaches, the population for which you are generalizing is ambiguous and often not well understood. The difficulty of understanding the nature of the population makes some kinds of statistical inferences difficult, improper, or meaningless.

Second, the various sampling techniques can use a variety of data collection methods. Convenience samples, for example, may be collected using paper and pencil, the Internet, observation, interviews, and so forth, making it a flexible sampling technique. Other sampling techniques (e.g., purposive sampling) may be less flexible, depending on the sample population one wishes to generalize to and how easy the population is to access. Additionally, although convenience sampling is the technique most commonly associated with online data collection methods, other sampling techniques are likely to use online data collection as well due to the convenience of this technological advancement. Thus, understanding potential issues that may arise when collecting data online is important across the various sampling techniques reviewed here.

Unsurprisingly, convenience sampling was used disproportionately in *Journal of Applied Psychology* articles in the past 5 years. Although these are only a "sample" of the samples used in organizational behavior research, from experience we have good reason to believe that this finding likely generalizes to other journals as well. As mentioned above, the reliance on convenience samples likely limits our ability to generalize. We suspect that for many researchers, the sample used may be an afterthought in many cases. Indeed, convenience samples may simply be a default due to their "convenience." In the following section, we identify considerations that should be made when deciding what sample to use by reviewing some key distinctions between sampling methods.

FACTORS TO CONSIDER WHEN CHOOSING A SAMPLING TECHNIQUE

Many considerations should guide the choice of a particular sampling technique when designing research. In choosing between various techniques, it is helpful for researchers to consider carefully the nature of their research question as well as the needs of the statistical methods used. Below we outline some of these considerations and how the sampling techniques discussed above may fit into them. **Table 2** summarizes these considerations.

Cost

Although cost should not be the main rationale for choosing a particular sampling technique, it would be unrealistic to ignore its importance for many researchers. Convenience sampling is often, though not always, less expensive than other sampling techniques. Other types of convenience samples may rely on student participation, which is often free; on personal connections and goodwill to entice respondents; and/or on an incentive raffle. Snowball samples are similar in terms of cost, relying on the goodwill of participants to identify additional participants. Purposive sampling is also typically low cost, relying on researchers to identify potential participants. Probability-based samples can be the most expensive, requiring access to enumerated lists of all participants who are members of the population.

Table 2 Key considerations when choosing samples

	Probability-based sampling	Stratified random sampling	Convenience sampling	Snowball sampling	Purposive sampling
Cost	Might need to purchase access to population list, cost of mailing to participants if email addresses are not available, incentives for participants	Same as for probability-based sampling	Cost of access to online panels, participant incentives if not a paid online panel	Possible participant incentives	Possible participant incentives
Accurate estimation of population descriptive statistics	Able to estimate the degree of accuracy	Able to estimate the degree of accuracy	Difficult to determine the degree of accuracy	Difficult to determine the degree of accuracy	Difficult to determine the degree of accuracy
Time to collect	Depends on whether email or home addresses are in the population list	Depends on whether email or home addresses are in the population list	Very fast for online panel samples; varies for other types of convenience samples	Somewhat slower, as it takes time for participants to identify likely respondents	Depends on researchers' ability to identify possible participants
Sufficient numbers of subgroup members	Can be a problem when sample size is small overall	Designed to collect sufficient number of subgroup members	Can be a problem	Can be a problem	Tends not to be a problem
Importance of generalizability	Generalizable to the population	Generalizable to the population	Unclear generalizability	Unclear generalizability	Unclear generalizability

Accurate Estimation of Descriptive Statistics

In some cases it is important to generate accurate estimates of means. For example, in political polling, the average amount of support that a candidate receives at a particular time is used to forecast an election. In addition, organizations may be interested in estimating the mean level of job satisfaction among employees at a particular time to compare with previous levels of satisfaction, or to compare with job satisfaction in other companies within the industry. In these cases, where it is important that descriptive statistics align with population values, PBS must be done. These types of research questions, however, are rare in organizational research.

Time to Collect

Convenience samples vary in how long it may take researchers to collect data. Student samples may be collected relatively quickly (e.g., during a class period) or over the course of a semester. Field samples may take more coordination and time, but once the planning and coordination are complete, the data collection may be fairly quick. Recently, online professional panels (e.g., MTurk, Prolific) have grown in popularity due to the rapid rate at which researchers can collect large quantities of data, sometimes within hours after the link to the survey has been posted. Purposive sampling may take relatively longer to identify respondents, as might snowball sampling, which requires participants to identify individual possible respondents. PBS depends on the nature of the access to participants, such as whether home addresses are used or whether email addresses are available.

Sufficient Numbers of Subgroup Members

For many research questions, we want to be able to compare groups with one another. For example, in measurement equivalence studies (Tay et al. 2015), researchers compare whether an instrument functions similarly across two or more groups. Purposive sampling can be used if researchers have access to enough individuals in subgroups. In many cases, however, it is a challenge to get enough members of a subgroup to complete surveys, especially when the base rate of the subgroups is relatively small in comparison to the majority group. In these situations, researchers can use stratified sampling to oversample subgroup members from the population; the advantage of doing so is that stratified random sampling still selects individuals from the population randomly, thus preserving the nonbiased nature of the design. If there is no enumeration of the population, or if the identity of subgroup members is not identified within that enumeration, then researchers must either use convenience sampling with screeners, relying on individuals to identify whether they belong to the subgroup of interest, or use snowball sampling, relying on individual participants to help identify other participants. In both cases, participant honesty is of concern if the identity of the target sample cannot be confirmed.

The Importance of Generalizability

When we think of sampling, we often think of generalizability. The importance of generalizability and the population of interest should dictate the choice of sampling approach. When we use PBS, we can determine with great precision the relation of the sample, and the statistics that we computed from that sample, to the population. In other sampling techniques, however, the relation of the sample to the population is murky at best. For example, with MTurk samples, suppose we screen for respondents who work at a job (outside of MTurk) for more than 20 hours a week. Does our sample of MTurk workers generalize to all workers who work 20 hours a week, or is our particular sample of MTurkers different from other workers who do not also participate on MTurk? Regarding this question, Highhouse (2009) makes a good point about generalizability. For much research, if you are attempting to understand basic psychological phenomena shared by all humans, then nearly any sample will suffice. Highhouse's research on basic judgment and decision-making is a good area for not worrying much about generalizability. In other areas, though, where we are interested in understanding the role of individual and organizational differences, concern for generalizability is much more salient. In these cases, individuals and organizations matter. With convenience, snowballing, and purposive samples, the population to which one sample is generalizing is unclear and concern for generalizability is more warranted.

WAYS TO IMPROVE THE QUALITY OF THE SAMPLE, ESPECIALLY SAMPLES COLLECTED ONLINE

Selecting a sample that is appropriate for one's research question is only half the sampling battle. Once an appropriate sample has been selected, researchers also want to ensure that the data being collected are of sufficient quality. Although the quality of responses is important for all sampling techniques and data collection methods, our attention is primarily on samples collected online (e.g., sampling platforms, social media) for several reasons. First, the remote nature of online sampling creates additional challenges that have raised legitimate concerns over whether online participants provide high-quality data or accurately represent their background and identity (Bernerth et al. 2021, Johnson 2005, Keith et al. 2017, Newman et al. 2021). Second, online samples have become the primary means of data collection—a phenomenon likely to continue. In our coded studies, 90.7% of all samples used online questionnaires to collect data. Third,

researchers may have less control over data collection for certain sampling techniques. For example, much of the research using PBS utilizes archival data. In such cases, the researcher has less control over ensuring data quality. Finally, research using certain data collection methods (e.g., interview, case study, observation) is less concerned with common data quality concerns such as insufficient effort responding (IER) or honest responding, as participant data collection is more interactive and the identities of these participants are typically confirmable.

When researchers discuss data quality within online samples, their primary concerns revolve around attention, honesty, and, more recently, bots (i.e., nonhuman participants). The anonymous nature and reward structures frequently present in online data collections (whether through a sampling platform such as MTurk or social media) create perverse incentives for higher instances of careless responding and dishonesty (e.g., misrepresenting one's eligibility). More recently, concerns have also been raised that bots may be responding to surveys (Kennedy et al. 2020, Moss et al. 2021, Newman et al. 2021, Storozuk et al. 2020). Each of these concerns has implications for the quality of our science; however, preventing careless responding, dishonesty, and bots typically requires the researcher to make different decisions during research design, data collection, and data cleaning stages. Below we provide an overview of each data quality concern along with current recommendations (for a summary of these recommendations, see **Table 3**).

Importantly, best practices for improving data quality are quickly evolving. For example, earlier research frequently recommended using a high approval rating when collecting data through MTurk (e.g., Fleischer et al. 2015, Peer et al. 2014). Later research, however, suggests that high approval ratings are insufficient to improve the quality of data collected on MTurk and may exacerbate the issue of nonnaïveté (Robinson et al. 2019). These recommendations are therefore based on the current state of our science; however, researchers should attempt to stay up to date as the science emerges.

Careless Responding

Careless responding, or IER, refers to instances in which participants exhibit "little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses" (Huang et al. 2012, p. 100). Estimates of IER in a typical study range broadly from 1% to 30%—though 10% to 15% is most likely for the typical survey (Curran 2016)—and may depend on factors such as sample characteristics or survey length (Bowling et al. 2016, 2021; Gibson & Bowling 2020; compare with Bowling et al. 2022). Even small amounts of IER, however, can increase total error by affecting observed correlations, particularly if the phenomenon of interest is high or low base rate (Credé 2010, DeSimone et al. 2017, Fleischer et al. 2015, Huang et al. 2015b, McGonagle et al. 2016). IER is likely to be common across all samples, regardless of the sample type or how it was obtained, but may be more likely in samples collected online, whether a sample of students, employees, or online panel participants (Huang et al. 2012, Meade & Craig 2012, Ran et al. 2015). Thus, it is essential that researchers be aware of and implement strategies to address IER both before and after data are collected. We provide an overview of the various means for screening IER; for a more thorough treatment of this topic, we refer readers to previous reviews and empirical investigations (e.g., Curran 2016, DeSimone et al. 2015, Meade & Craig 2012).

For our purposes, we provide a brief overview of seven categories of IER screening (for a summary of these approaches, see **Table 4**): attention checks, Mahalanobis distance, long string, response time, consistency indices, response coherence, and self-reported effort. Critically, best practices recommend using more than one technique, depending on one's research design and study needs (Curran 2016, DeSimone et al. 2015, Dunn et al. 2018). IER indices correlate but often do not converge; each approach detects different forms of IER and may have different strengths and weaknesses (Curran 2016, DeSimone & Harms 2018, Huang et al. 2012, Meade & Craig

Table 3 Summary of recommendations for improving data quality

Topic area	Specific recommendations
General	1. Decide a priori how you will screen for careless responding, dishonesty, and bots. 2. Transparently report any prescreening or data cleaning efforts in the Method section. 3. Analyze data before and after screening out participants; report any differences. 4. Avoid overzealous data cleaning. Use conservative cutoffs (see **Table 4**), and consider the nature of your research design when setting cutoffs. 5. Provide compensation that is proportional to the time and effort required. 6. Remain current on best practices for improving data quality.
Careless responding	1. Use multiple indices of IER. 2. Use attention check items that have a similar stem and length as other scale items instead of generic items (e.g., "Please select 'strongly disagree'"). 3. Use instructional manipulation checks when important for one's research design (e.g., experimental manipulations); otherwise, use sparingly. 4. Use bogus/infrequency items that are less likely to be interpreted figuratively or prompt social desirability/impression management. Consider pilot testing items. 5. Embed attention checks throughout the survey. 6. Consider using statistical approaches such as Mahalanobis distance, psychometric antonyms/synonyms, and individual reliability approaches when appropriate for the research design and scales used. 7. Always collect timing information for each page of the survey. 8. Where appropriate, include open-ended items to gauge motivation with response coherence and quality. 9. Use self-reported effort/attention and long string only in conjunction with other indices. 10. When possible, shorten the length of surveys, make them interesting, and design surveys with the participant in mind.
Honesty	1. Use a prescreen (e.g., a smaller separate survey) to identify participants of the desired target population. It should not be obvious that the survey is to determine eligibility for a separate survey or what the eligibility requirements might be. 2. If a prescreen is used to identify a target population, repeat items assessing targeted characteristics. 3. Embed questions asking similar things at separate points in the survey to screen for inconsistencies (e.g., "How old are you?"/"What year were you born?"). Ensure that inconsistencies are not due to misinterpreting one or both questions. 4. Use available tools to screen out VPS participants and check IP addresses. 5. Consider sampling techniques or recruitment practices that verify the identity of participants when the demographic characteristics are critical to your research questions.
Bots	1. Consider the level of risk with a given sample. Student samples or direct recruitment are low risk, samples recruited from social media are high risk, and samples collected through online platforms are moderate risk. 2. Include CAPTCHA and honey pot items, and select "Prevent multiple submissions" when designing surveys. 3. During and after data collection, examine click counts, email addresses (if collected), page timing information, and responses to open-ended items to identify nonhuman respondents (bots). 4. Include language in recruitment statement and consent form about refusing payment to bots. 5. Avoid autopaying participants. 6. If using an online platform, notify the platform about suspicious activity.

Abbreviations: IER, insufficient effort responding; VPS, virtual private server.

2012). Also, there are different cutoff recommendations for many of these indices, requiring some amount of researcher judgment. We echo other researchers who recommend deciding how IER indices will be used a priori, transparently reporting this information in one's Method section, and transparently reporting any differences in results between screened and unscreened data (Curran 2016, DeSimone & Harms 2018, Keith et al. 2017).

Table 4 Means of detecting IER

Method	Definition	Flag criteria	Limitations
Attention checks	Items or instructions embedded in a survey attempting to catch people who are not reading instructions or items		
Instructed items	Items embedded in a survey attempting to catch people who are not reading instructions or items Example: "Please respond 'strongly agree' to this item"	Participants who fail to answer >50% of items as instructed	Participants can search out attention check items that are obvious (e.g., different stem, different length) and randomly click through the remainder of the survey items.
Instructional manipulation checks	Asks participants to recall something from instructions	Participants who fail to recall information from instructions	Potentially confounds reading instructions with reading items participants are responding to May be easily identifiable by experienced survey takers
Bogus/infrequency items	Asks participant agreement on something improbable Examples: "I was born on February 30," "I have never been angry"	Participants who answer >50% of items with anything other than "strongly disagree" or "disagree" to items	Items may be interpreted figuratively rather than literally. Potentially confounds IER and faking/impression management
Mahalanobis *D*	An outlier statistic that uses a chi-square test to detect multivariate outliers; estimate of the multivariate distance between participant's score on items and sample mean scores on items Assumes that an extreme deviation from normative response pattern indicates lack of attention Squared value is distributed as a chi-square value.	Participants whose D^2 values put them in the top 5% of the chi-square distribution	Not effective for detecting careless responding when careless responding follows a normal distribution for all items Requires measures with a large number of items
Long string	Calculates number of consecutive identical responses on a Likert scale (e.g., responding "agree" 20 times in a row) Can measure average or maximum long string	6–14 invariant responses, but depends on the nature of the scale	Identifies only one form of IER; cannot detect random responding Not good for use on a homogeneous scale Cannot be used if items are randomized within a scale or set of scales
Response time	Time it takes to complete items Can be normed within sample or use a rule of thumb, such as 2 s per item	Response time of <2 s per item Participants who are much faster than a certain percentage of the sample	Some variability can be expected with differences in reading speed, familiarity with scale, decision-making speed, language fluency, etc. The cutoff of 2 s per item is somewhat arbitrary. Does not capture those who are not speeding through but are otherwise distracted

(*Continued*)

Table 4 (*Continued*)

Method	Definition	Flag criteria	Limitations
Consistency indices	Assumes that there should be some stability in the response patterns for items of the same dimension		
Even–odd consistency	Divide scale into half by odd/even items. Correlate scores for each half. The closer to zero, the less consistent Spearman–Brown split-half formula	Individual reliability <0.30	Performs poorly when participant is only occasionally responding carelessly
Response consistency	Gives participants the same items within the same session and examines correlations between responses The closer to zero, the less consistent	Response consistencies <0.25	Need 30 repeated items within survey to get a relatively stable response Can result in additional fatigue and irritation for participants Not good for measures that are likely to have high within-person variance (e.g., affect) or shorter surveys
Semantic antonyms/ synonyms	Examines within-person correlation between (dis)similar pairs of items Can also use the same item as with response consistency Should be separated in the survey	No currently agreed upon cutoff; should not require perfect agreement	Requires human judgment to determine similarity or dissimilarity of item content Not good for measures that are likely to have high within-person variance (e.g., affect) or shorter surveys
Psychometric synonyms/ antonyms	Within-person correlation between pairs of items that have the strongest positive/negative within-person correlation	0.60/−0.60 cutoff for identifying pairs <0.22 for flagging psychometric synonyms >0.03 for flagging psychometric antonyms	May not be able to statistically identify item pairs or may have too few; psychometric antonyms more effective if have both positively and negatively worded items in survey Not good for measures that are likely to have high within-person variance (e.g., affect) or shorter surveys
Response coherence	Gives participants open-ended items and examines whether the response makes sense	NA	May result in unwanted self-selection of participants Will flag bots and nonnative speakers, not necessarily careless responding
Self-reported effort/attention	Asks participants how much effort they put into responding to the survey Can be a single item or multiple items	Self-reporting lack of attention or effort	Transparent and susceptible to dishonesty

Abbreviations: IER, insufficient effort responding; NA, not applicable.

Attention checks. Attention checks are items or instructions embedded into a survey that attempt to gauge whether participants are reading and comprehending instructions or survey items. There are several different types of attention checks that generally fall into two categories: instructed or bogus/infrequency.

Instructed attention checks either prompt participants to respond to a survey item in a particular way (instructed items) or ask participants to recall something from the instructions (instructional manipulation check) (Huang et al. 2012, Meade & Craig 2012, Oppenheimer et al. 2009). Instructed items include embedded items such as "Please respond 'strongly disagree' to this item," "Please leave this item blank," or "If you are paying attention, respond 'agree' to this item" (DeSimone et al. 2015, Meade & Craig 2012). Instructional manipulation checks may embed a directive within a larger instructional paragraph. Participants are often asked to ignore the rest of the instructions and demonstrate that they are paying attention by ignoring the set of items and proceeding without selecting an item, by selecting a set of items, or by entering something into an open text box (Oppenheimer et al. 2009). In either case, participants who fail to follow instructions would be presumed to not be reading or comprehending the instruction or survey item.

Previous research has validated the use of both types of instructed attention checks (DeSimone et al. 2015, Kung et al. 2018, Oppenheimer et al. 2009). Kung et al. (2018), for example, found that instructed items and instructional manipulation checks did not affect the scale means or the way the scale was interpreted by participants. Although instructed attention checks may be useful for detecting some degree of IER, we suspect that their efficacy may wane over time among more-experienced survey takers. Hauser & Schwarz (2016) suggest from their findings that MTurk participants have learned over time how to evade instructional manipulation checks. Indeed, the second author of this review spent some time as a participant for Prolific and found instructional manipulation checks in most research studies. The formulaic nature of both types of attention checks also makes them fairly easy to identify. For example, researchers frequently use generic instructed items such as "Please respond 'strongly disagree' to this item" (Meade & Craig 2012). Experienced survey takers, however, may be quickly able to identify these items on a page and randomly click through the remaining items (Barends & de Vries 2019, Hauser & Schwarz 2016). Thus, the combination of common and easily identifiable weakens these items' effectiveness with more-seasoned survey takers (Curran 2016, Hauser & Schwarz 2016).

To increase the efficacy of instructed items, we recommend using items that have a similar stem and length as the other scale items in lieu of more generic items. In addition to these concerns about efficacy, we caution that asking participants to recall something from the instructions may have limited construct validity, as this type of attention check may confound participants' motivation to read instructions and motivation to read survey items. In other words, participants may be accustomed to skipping over generic instructions such as "Please tell us more about yourself" or "Please rate the extent to which you agree or disagree with each statement." Thus, unless the instructions are critical to the research design (e.g., part of an experimental manipulation), we advise caution when using instructional manipulation checks.

Bogus/infrequency attention checks use survey items that are improbable (e.g., "I was born on February 30" or "I have never been angry"; Huang et al. 2015a, Meade & Craig 2012). Participants who agree to improbable survey items are flagged. The main concern with bogus items is that they may lack construct validity; such items may be interpreted figuratively or confound IER with faking/impression management (Curran & Hauser 2019, Huang et al. 2015a). For example, Curran & Hauser (2019) found that some participants who were paid biweekly agreed with the item "I get paid by leprechauns biweekly," suggesting that they may have chosen to pay attention only to the biweekly portion or that they interpreted leprechauns figuratively rather than literally. Many infrequency items are also used to detect socially desirable responding, raising the question of whether infrequency items are appropriate for detecting IER (Huang et al. 2012, 2015a). For example, a participant may respond to improbable items such as "I have never been absent from work" in the affirmative due to impression management rather than lack of attention. Notably, Meade & Craig (2012) provided evidence that infrequency items loaded onto the

same factor as other IER inconsistency indices and were distinguishable from socially desirable responding. Huang et al. (2015a) provided further validity evidence for infrequency items and found that such items do not result in more-negative participant reactions. For researchers using bogus/infrequency items, we recommend pilot-testing these items and designing items that are less likely to be interpreted figuratively or prompt social desirability.

Regardless of the type of attention check being used, we echo other researchers in recommending conservative cutoffs (e.g., failing 50%) rather than a zero-tolerance approach (Curran 2016, Curran & Hauser 2019; compare with Kam & Chan 2018, Kim et al. 2018). Removing a participant who misses one attention check runs the risk of eliminating a participant who either (*a*) misunderstood the item or (*b*) was otherwise attentive. It is also reasonable to expect attention to wane, especially for surveys that are lengthy. Thus, it is important to attempt to limit the length of surveys when possible and embed attention checks throughout the survey to detect IER that is inconsistent (Bowling et al. 2021).

Mahalanobis distance. Mahalanobis distance (Mahalanobis *D*) is a form of multivariate outlier analysis that provides an estimate of the multivariate distance between a participant's response to survey items and the response of the rest of the sample (Curran 2016, Mahalanobis 1936). The assumption is that an extreme deviation from a normative response pattern indicates a lack of attention. Although there are no established cutoffs for this index, DeSimone et al. (2015) suggested that the squared value of Mahalanobis *D* is a chi-square distribution and that researchers could flag participants who are in the top 5% of the distribution.

Unlike some of the other IER indices discussed here, Mahalanobis *D* requires more effortful calculations and a large number of items (Meade & Craig 2012). In order to be flagged as IER, this technique also requires careless responses to diverge from a normal distribution. For these reasons, Mahalanobis *D* may not be appropriate for shorter surveys or careless responding that follows a normal distribution.

Long string. One form of IER is when participants respond to survey items with consecutive identical responses (e.g., responding "strongly agree" for an entire scale; Costa & McCrae 2008, Johnson 2005). To calculate long string, researchers can take a range of approaches, including averaging the invariant responses, examining the maximum long string across scales, and other similar calculations (Curran 2016). Cutoffs of 6–14 invariant responses have been proposed (Costa & McCrae 2008); however, long string requires researchers to use their judgment as to what is a reasonable cutoff for the nature of the scales being used. For example, it does not make sense to set a cutoff of 10 long string responses if a scale has only 5 items. Additionally, if items are homogeneous it is not unreasonable for participants to select "agree" or "disagree" for several items in a row (Curran 2016, DeSimone et al. 2015). Long string also cannot be used if items are randomized within a scale or set of scales.

It is unlikely that long string will be useful except in the most extreme cases of carelessness or lack of effort. Most participants who are familiar with surveys will know to vary their responses at least a little to avoid being detected. For participants who vary their responses even slightly, long string will not be able to detect such instances of random responding—an approach that is likely to be more common. Thus, examining one's data only for long string responses is insufficient for detecting most cases of IER. When using long string, we urge researchers to consider the nature of their scales and use a conservative cutoff decided a priori to flag participants.

Response time. How long participants spend on survey items is a somewhat intuitive measure of participant effort. As Curran (2016, p. 6) aptly put it, "response time does appear to be one of

the hardest metrics of [IER] to fool. . .because of one simple fact: a presumed key underlying the motive of [IER] responders is finishing the assessment as quickly as possible." Response time can be examined either by using sample norms to identify outliers (i.e., participants who spent too little or sometimes too much time on a survey) or by looking at the response time per item.

As with the other IER indices, response time—though intuitive—has certain limitations. Although a cutoff of 2 seconds per item (Huang et al. 2012) has been widely adopted, there is reasonable variation in reading speed, decision-making speed, literacy, and familiarity with the scale. Due to nonnaïveté, MTurk participants may be more familiar with commonly used scales, resulting in faster-than-average response times (Chandler et al. 2014, Keith et al. 2017). It is also not clear what to do when participants spend much longer than the average participant on a survey (Curran 2016). A long survey time could indicate a range of things, including taking the survey while distracted by one's environment, taking multiple surveys at once, or simply taking a break in the middle of a survey to recharge. Only some of these potentialities are problematic, but it can be difficult to determine which might be occurring without examining other IER indices.

Despite these potential limitations, response time has been found to be an effective means of detecting IER (Bowling et al. 2021, Huang et al. 2012, Wood et al. 2017). Even if the researcher does not ultimately use response time, we highly recommend collecting response times on each page of the survey, as doing so is noninvasive, is easy to implement, and cannot be done retroactively after the data are collected. Most survey platforms (e.g., Qualtrics) automatically provide timing information for the entire survey; however, this is less ideal than collecting timing for individual survey pages, for two main reasons. First, IER may not be consistent across a survey; participants may respond carefully at the beginning of a survey but begin responding carelessly as fatigue sets in (Bowling et al. 2021). Second, collecting timing information for the survey as a whole may obfuscate IER if participants spend disproportionately longer on one survey page and rush through the other pages. For example, a participant may pause to get a drink of water before returning to randomly responding to survey items.

Consistency indices. Several methods exist to examine whether a participant's pattern of responses is internally consistent. Each assumes that participants should have some stability in their response patterns for similar items or items of the same dimension. We discuss a few of these methods below and refer readers to Curran (2016) for a more in-depth account of consistency indices, including ones not discussed here.

Even–odd consistency divides a scale in half by odd/even items. The researcher then takes the average scores from each half and finds the correlation between them. The Spearman–Brown prophesy formula is also typically used to correct for scale length (DeSimone et al. 2015, Johnson 2005). A low individual reliability (e.g., below 0.30; Johnson 2005) would indicate that the participant was inconsistent in their response pattern within the scale. Curran (2016) pointed out that computational advances have negated the need to simplify calculations by taking one pair of items to split a scale in half; there are now better ways to examine individual reliability. Curran went on to recommend resampled individual reliability, a technique that uses resampling of item pairs to reduce random error. This technique, however, requires more research in its application to IER.

Another method of examining internal consistency is to give participants the same items within a survey and examine the correlations between the responses. This technique, known as response consistency, assumes that individuals responding attentively will provide similar responses within a survey (Wood et al. 2017). Notably, past research recommends using 30 repeated items in order to get a reliable response consistency index (Wood et al. 2017). As a result, participants may become more fatigued or even irritated if they start to recognize that some items are being asked more than once.

The last set of consistency indices that we discuss here includes semantic and psychometric synonyms and antonyms. These techniques use pairs of either similar items (synonyms) or dissimilar items (antonyms) to examine within-person correlations between sets of items (DeSimone et al. 2015). Semantic synonym or antonym pairs are identified using human judgment of the similarity or dissimilarity of the item content. There is no currently agreed upon cutoff for identifying IER with semantic synonyms or antonyms; however, researchers should not expect perfect agreement (DeSimone et al. 2015). Psychometric synonym or antonym pairs are identified by examining correlations between items in the full sample (between-person correlations). Past research has recommended a 0.60/−0.60 cutoff for identifying pairs of psychometric synonyms or antonyms (Curran 2016, DeSimone et al. 2015, Meade & Craig 2012). The cutoff for identifying IER varies, but previous research has used a <0.22 cutoff for psychometric synonyms and a >0.03 cutoff for flagging psychometric antonyms (Huang et al. 2012, Johnson 2005, Meade & Craig 2012). Each technique is most useful with scales that have both positively and negatively worded survey items—because it is not guaranteed that psychometric synonym and antonym pairs will be identified in any given sample—and should not be used for measures that may have high within-person variance (e.g., affect) (DeSimone et al. 2015).

Response coherence. Items on a Likert scale are the modal means of collecting data; however, these measures are more prone to careless responding. Conversely, open-ended items are less easy to respond to carelessly—at least, not in a way that is inconspicuous to researchers. Having required open-ended items in a survey may be beneficial to researchers for two main reasons. First, requiring participants to respond to a qualitative item toward the beginning of a study may cause participants who would prefer to put little effort into a survey to opt out of the study. In other words, some self-selection may take place that results in a more motivated sample of participants. Second, if participants are responding carelessly or are nonhuman bots, it will be very clear on the basis of the coherence and quality of their response to an open-ended item (Dupuis et al. 2019, Storozuk et al. 2020).

There are, however, a few potential limitations to this approach. The selection that takes place from participants being screened out or opting out of surveys that require open-ended responses may result in a sample that differs from the sample population or the general population on certain individual differences (e.g., conscientiousness, agreeableness) or other characteristics (Bowling et al. 2016, Dunn et al. 2018). It is also possible that response coherence is a better screening tool for bots or nonnative English speakers than for careless responding. That is, participants may put effort into their qualitative responses but respond carelessly to scale items.

Despite these potential limitations, we encourage researchers to use open-ended items where appropriate for their research design and goals. Participants who decide to remain in a study that requires qualitative responses may find the task more enjoyable than simply clicking bubbles on a screen. Anecdotally, we have received comments from participants completing idea generation tasks such as "It was fun," "The study was interesting and enjoyable to do," "Definitely a fun activity!" and "Interesting and thought provoking." This feedback also highlights the importance of designing studies with participants in mind more generally. To help ensure that participants do not feel exploited with open-ended response requirements, we recommend increasing pay (if applicable) to be proportional to the additional time and effort required.

Self-reported effort/attention. The final means of detecting IER simply asks participants how much effort they put into responding to the survey or their level of attention. Participants who respond that they put in less effort or did not pay attention may be flagged. Importantly, self-reports are only useful for excluding participants who admit to responding carelessly and should

not be tied to compensation or other incentives to encourage more honest responding. Although there is some support for the efficacy of self-report measures of effort (Huang et al. 2012, Meade & Craig 2012), these measures are susceptible to faking and impression management (DeSimone et al. 2015, Meade & Craig 2012). Researchers often combat these issues by letting participants know that their answer will not affect compensation; however, this requires participants to read the instructions and trust the researchers. We view this technique as relatively limited in its ability to identify IER and recommend using it only in conjunction with other indices.

Honesty

Honesty—or, in this context, dishonesty—involves the intentional misrepresentation of one's demographic characteristics, attitudes, behaviors, cognitions, and so forth. For example, MTurk participants may misrepresent their IP addresses to qualify for a study that is intended only for US participants (Dennis et al. 2020) or demographic characteristics to qualify for a study intended for a specific population, such as people with depression, LGBTQ+ people, or working adults (Bernerth et al. 2021, Chandler & Paolacci 2017, Kan & Drummey 2018, MacInnis et al. 2020). Dishonesty poses an obstacle for organizational researchers who are interested in recruiting a sample from a specific population of interest such as employed individuals or one that is otherwise part of a particular demographic.

The issue of dishonesty also appears to be prevalent and can manifest in different ways. The first manifestation involves dishonesty around one's demographic characteristics. For example, Kan & Drummey (2018) examined whether participants on MTurk would misrepresent demographic characteristics across two studies. In their first study, they found that 55.8% of participants misrepresented their color-blind status, with 22.8% reporting that they had red/blue color blindness—a type of color blindness that does not exist. In their second study, they found inconsistencies in self-reported demographic characteristics among participants completing studies at different time points. Dishonesty ranged from 6.6% to 38.2% depending on the demographic characteristic (age, 22.6% inconsistent; education, 31.3% inconsistent; gender, 6.6% inconsistent; income, 38.2% inconsistent; family status, 14.8% inconsistent). Similarly, MacInnis et al. (2020) found that, depending on the honesty metric used, between 2.2% and 28% of participants ($N = 4{,}128$) appeared dishonest. Between a prescreen and a main survey, 28% reported different levels of hiring authority in their organization and 3.5% reported different ages. Additionally, 2.2% of participants reported that they had experience with a fictional organizational system. MacInnis and colleagues also found that those who misrepresented their age were lower in honesty/humility than those who did not misrepresent their age ($d = 0.50$); however, the same was not found for those who misrepresented their level of hiring authority.

Another common form of dishonesty is the use of virtual private servers (VPSs) to mask one's IP address to misrepresent location (Kennedy et al. 2020). Researchers frequently limit their samples to US participants; however, participants in other countries may override this imposed qualification using VPSs. Kennedy et al. (2020) examined IP addresses for surveys conducted between 2013 and 2018 and found that between 15% and 20% of respondents were VPS users. Moreover, these VPS users appeared to provide poorer-quality data, with 23.9% being flagged by at least one quality check compared with 2.8% of non-VPS users.

Notably, it is often difficult to determine whether misrepresentation is intentional or simply the result of a lack of attention or understanding the instructions/question. For example, some research uses bogus items (e.g., "I have experience with the BTE organizational system"; MacInnis et al. 2020) that may tap into either an intentional misrepresentation of one's experiences or IER. What is clear, however, is that dishonesty can and does happen, and instances of dishonesty can

result in misleading conclusions or generalizations to population of interest (Chandler et al. 2020, Kan & Drummey 2018).

Three main recommendations have been made to address honesty. First, researchers attempting to target a particular population should use nonexplicit prescreens. Kan & Drummey (2018) used explicit eligibility requirements and found high rates of misrepresentation as a result. In contrast, MacInnis et al. (2020) did not use explicit prescreens and found high rates of dishonesty only for hiring authority. These results suggest that some dishonesty can be prevented by obscuring the intention of a prescreen; however, for rare or specific populations (e.g., LGBTQ+), participants may still attempt to guess what researchers are looking for. We recommend repeating items assessing eligibility in the main study to identify inconsistencies. Researchers may also consider the second recommendation for detecting dishonesty: If hypotheses or research questions rely on specific characteristics or constructs, researchers may want to embed separate questions within the survey to screen for inconsistent reporting. For example, if the survey is intended for working parents, researchers may want to embed one question at the beginning of the survey asking how many children the person has currently living at home and their ages. Then, later in the survey or at a second time point (if using a time-separated design), the researcher could ask what year each of their children was born. Inconsistencies may be indicative of inaccurately reporting one's parental status. When using this recommendation, the researcher should ensure that the embedded items provide a clear indication of dishonesty rather than simply reflecting a different interpretation of the question. Third, researchers should use available tools to prevent responses from VPS users or suspicious geocode locations. Online sampling platforms (e.g., CloudResearch, Prolific) typically provide tools for screening out VPS participants and cross-checking IP addresses. Researchers may also want to consider available R packages for screening IP addresses (Kennedy et al. 2020, Waggoner et al. 2019). Finally, we add the suggestion to include manipulation checks and/or open-ended questions within surveys, as previous research (e.g., Dennis et al. 2020) has found that VPS participants are more likely to fail manipulation checks and 81–91% of VPS responses provided incoherent or nonsensical responses to open-ended questions.

The above recommendations may be sufficient when the characteristics of one's sample are not critical for one's research question. If, however, researchers are attempting to generalize to a very specific population, we recommend using sampling techniques or recruitment procedures that allow researchers to verify the desired characteristics of their sample in some way. Examples are direct recruitment of the desired population (e.g., use of directories), as is sometimes done in PBS and purposive sampling, and the inclusion of an in-person consent procedure.

Bots

Bots are nonhuman respondents that are programmed to automatically respond to surveys (Kennedy et al. 2020, Storozuk et al. 2020). Using autofill or other widely available software, bots can quickly submit hundreds or even thousands of responses to a survey posted online within hours, creating a huge problem for researchers. Indeed, concerns over nonhuman responses have grown substantially in recent years alongside the expanding availability of programs available to even minimally experienced hackers (Aguinis et al. 2021, Newman et al. 2021, Storozuk et al. 2020). Indeed, the pervasiveness of data generated by bots is currently unknown and difficult to quantify, as some of the data originally suspected of being bot-generated were later determined to be human-generated (i.e., fraudulent VPS respondents or humans located outside the USA using VPSs to take surveys intended for US participants; Kennedy et al. 2020). In other words, in most cases the bot crisis may be more of a dishonesty crisis. When bots do strike, however, it can be a huge headache for researchers.

So, what do we do about bots? To start, researchers should consider their level of risk concerning bots and, where there is risk, take measures to prevent bots prior to collecting data. Bots are not interested in hacking a survey for course credit; however, they may be attracted to well-paying studies posted online. Many online platforms such as CloudResearch and Prolific have measures in place to block bots and the same or suspicious geocode locations, thereby reducing the likelihood of bots on these platforms (Bradley 2018, Moss & Litman 2018). Anecdotally, even with these measures, bots appear to be able to bypass some of these measures. Conversely, social media sites such as Facebook and Twitter may be especially susceptible to bot activity, as bots frequently use these websites to identify high-paying research studies (Pozzar et al. 2020). For this reason, researchers may want to avoid posting a public link to the survey and instead have participants contact researchers for a link to the survey. Alternatively, if researchers are interested in recruiting a particular population, it may be safer to recruit within a private Facebook group geared toward that population of interest or use a different recruitment method such as snowball sampling. Another option is to have participants verify their identity; however, this option must be weighed against participant rights to privacy (Godinho et al. 2020, Teitcher et al. 2015).

When designing a survey, researchers can also take measures to prevent bots by including a CAPTCHA, honey pot items (i.e., items that are invisible to human participants but visible to bots), and selecting "prevent multiple submissions" (formerly "prevent ballot box stuffing") when using survey platforms such as Qualtrics (Simone 2019, Storozuk et al. 2020). Many of the techniques for identifying IER may also be used to identify bots. Dupuis et al. (2019) compared seven different IER indices' ability to detect bots responding randomly and found that response coherence, Mahalanobis *D*, and person-total correlation were the most effective. Of course, it can be difficult to determine whether participants screened using IER indices are humans responding carelessly or nonhumans, and the technique used assumes that bots provide only random responses. Buchanan & Scofield (2018) used a bot constructed through Python and found that bots using autofill functions are likely to have improbably low click counts (i.e., fewer clicks than the number of items on a page), spend less time on a page, fail manipulation checks, and use more scale items (e.g., four or more), suggesting more random responding. Storozuk et al. (2020) noted that if a study has been infiltrated by bots, researchers may notice a large uptick in the number of participants, multiple responses with the same start and end times, suspicious email addresses (e.g., temporary email addresses, series of email addresses following a similar pattern), and nonsensical or similar responses to open-ended questions (e.g., strings of random letters, irrelevant responses such as "NICE," or multiple responses with the same answers). Taken together, if measures to prevent bots fail, bots can typically be identified by examining click counts, timing information, and responses to open-ended items and attention checks.

Bots and insufficient effort responders more generally can be a drain on valuable monetary resources unless researchers take these precautions. To avoid ethical problems surrounding transparency or trouble with institutional review boards, we also recommend including language in recruitment statements and consent forms about withholding payment for participants identified as bots. Researchers should also screen their data prior to paying participants to avoid paying large sums of money to bots. Finally, researchers are encouraged to notify the online platforms they are using about suspicious activity; these online platforms may be able to block these individuals or find new ways of preventing bots from infiltrating other research studies.

Unfortunately, bots pose an ongoing challenge, as they tend to become more sophisticated over time. For example, bots created by sophisticated programmers can be (and are) programmed to spend an appropriate amount of time on a particular page, provide responses that look like attentive responses, and even bypass CAPTCHA or honey pot items (Griffin et al. 2021, Storozuk

et al. 2020, Teitcher et al. 2015). Like IER, multiple safeguards are likely needed to catch more sophisticated bots. This unfortunate reality will also require researchers to be vigilant in keeping up with the latest advances attempting to stay one step ahead of bots.

Final Thoughts on Data Quality

In this section, we have described best practices for improving sample quality by identifying participants who are providing insufficient effort, are responding dishonestly, or may not even be human. These recommendations are tailored to online data collections, which, in our coding of the last 5 years of *Journal of Applied Psychology* articles, are slightly more than 90% of all samples. Data collections that require more person-to-person interaction, such as observations, interviews, and paper administration of surveys, are still susceptible to issues of insufficient responding and dishonesty, and many of the techniques that were created to improve online samples can be used for these other modalities, though often with some adaptation needed.

CONCLUSIONS

In this article, we have reviewed different sampling techniques, including several, such as purposive sampling and PBS, that are infrequently used in organizational research. In addition, we have reviewed current best practices on how to improve sample quality through a variety of screening and survey design techniques. In this final section, we reflect on the nature of sampling in organizational research and describe several concerns that we feel all organizational researchers should consider. Before offering our conclusions, we note that we coded articles only from the *Journal of Applied Psychology*; we suspect that our conclusions would be similar for other journals in organizational behavior and industrial–organizational psychology that focus on publishing quantitative research.

Concern for Lack of Methodological Diversity

The ease and availability plus the relatively low cost of online panel designs have brought about a lack of methodological diversity within organizational research and similar fields (e.g., Anderson et al. 2019). As shown in our review, the percentage of articles that use online panel studies is quite high, so data collection techniques that cannot be administered via online survey platforms are now less likely to occur. For example, observation of real work behavior, either in the laboratory or in naturalistic settings, is viewed as an important part of social science research, though the percentage of studies in our review that used observation is low (only 1.4% of samples). The relative ease of online panel sampling has shaped the types and forms of questions that we are more likely to ask now (Anderson et al. 2019).

Concern for Lack of Participant and Organizational Diversity

The overreliance on online panel samples specifically and convenience samples in general means that the participants who form the basis of our science likely do not mirror the populations as a whole to whom we wish to generalize our findings. As Bergman & Jean (2016) noted, if we consider the field of organizational research as a whole, we oversample certain types of workers and undersample others. Although we tend not to conduct research that makes strict generalization to a population value important (as might be the case in medicine or political science), by limiting our potential pool of participants to those who are in online panel samples or in convenient samples, we limit our ability to better understand boundary conditions to our findings.

Concern for Participant Quality Is Important

The focus that online panel surveys have brought to light is the importance of the quality of participant engagement with the survey. As reviewed here, researchers have made significant strides in identifying which participants are taking the survey seriously and which participants are responding haphazardly or inattentively, or in fact are not even humans. As technology continues to advance, and as the online participant economy becomes more sophisticated, vigilance in understanding and screening for participant quality will continue to be important and will require researchers to devote significant amounts of time understanding technology. Staying ahead of participant quality seems a bit like the game Whac-A-Mole: Once you learn how to screen for a new way of cheating, sophisticated participants will find a way to circumvent the new screening technique, and so on and so on (like an arm's race). It is important for researchers to keep pursuing these technological advances, though it is also important to take a step back and think about how to help motivate participants to care about providing high-quality and accurate data.

Choose Your Sample to Best Answer Your Research Question

Research questions should drive research programs, though we are not so naïve as to believe that sometimes access to a particular sample may provide the impetus for a research project. In general, though, sampling approaches should be chosen to collect the most appropriate data that can be used to answer research questions.

Importance of Reporting Key Details About Your Samples

In our review of existing practices, it is clear that authors varied in terms of the amount of detail they reported about their sampling. It is important for researchers to provide key details about the type of procedures they used to obtain their samples as well as to describe the key characteristics of each sample. We like the practice of several journals that additionally require contextual information about each particular sample. Many researchers have a "need to know" attitude about sample reporting, in that they report basic information (age, sex, and gender breakdowns) and hold back any other information that they do not deem important for their findings. We understand the importance of brevity, though more information can be helpful for future researchers.

Final Thoughts

Advances in technology have provided significant opportunities for organizational researchers to collect data in efficient ways that were not possible even 20 years ago. Online data collections have greatly changed the ways in which we collect our samples, although, as observed in this article, new and significant challenges abound. It is important for researchers to consider the impact of their sampling strategies as well as ways to improve the quality of their data collections.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Aguinis H, Villamor I, Ramani RS. 2021. MTurk research: review and recommendations. *J. Manag.* 47(4):823–37

Anderson CA, Allen JJ, Plante C, Quigley-McBride A, Lovett A, Rokkum JN. 2019. The MTurkification of social and personality psychology. *Personal. Soc. Psychol. Bull.* 45(6):842–50

Barends AJ, de Vries RE. 2019. Noncompliant responding: comparing exclusion criteria in MTurk personality research to improve data quality. *Personal. Individ. Differ.* 143:84–89

Bergman M, Jean V. 2016. Where have all the "workers" gone? A critical analysis of the unrepresentativeness of our samples relative to the labor market in the industrial–organizational psychology literature. *Ind. Organ. Psychol.* 9(1):84–113

Bernerth JB, Aguinis H, Taylor EC. 2021. Detecting false identities: a solution to improve web-based surveys and research on leadership and health/well-being. *J. Occup. Health Psychol.* 26(6):564–81

Bowling NA, Gibson AM, DeSimone JA. 2022. Stop with the questions already! Does data quality suffer for scales positioned near the end of a lengthy questionnaire? *J. Bus. Psychol.* 37:1099–116

Bowling NA, Gibson AM, Houpt JW, Brower CK. 2021. Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organ. Res. Methods* 24(4):718–38

Bowling NA, Huang JL, Bragg CB, Khazon S, Liu M, Blackmore CE. 2016. Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *J. Personal. Soc. Psychol.* 111(2):218–29

Bradley P. 2018. Bots and data quality on crowdsourcing platforms. *Prolific Blog*, Aug. 10. **https://blog.prolific.co/bots-and-data-quality-on-crowdsourcing-platforms**

Buchanan EM, Scofield JE. 2018. Methods to detect low quality data and its implication for psychological research. *Behav. Res. Methods* 50(6):2586–96

Chandler J, Mueller P, Paolacci G. 2014. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav. Res. Methods* 46(1):112–30

Chandler J, Paolacci G. 2017. Lie for a dime: when most prescreening responses are honest but most study participants are impostors. *Soc. Psychol. Personal. Sci.* 8(5):500–8

Chandler J, Sisso I, Shapiro D. 2020. Participant carelessness and fraud: consequences for clinical research and potential solutions. *J. Abnorm. Psychol.* 129(1):49–55

Costa PT, McCrae RR. 2008. The revised NEO personality inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment. Personality Measurement and Testing*, ed. DH Saklofske, pp. 179–98. Thousand Oaks, CA: SAGE

Credé M. 2010. Random responding as a threat to the validity of effect size estimates in correlational research. *Educ. Psychol. Meas.* 70(4):596–612

Curran PG. 2016. Methods for the detection of carelessly invalid responses in survey data. *J. Exp. Soc. Psychol.* 66:4–19

Curran PG, Hauser KA. 2019. I'm paid biweekly, just not by leprechauns: evaluating valid-but-incorrect response rates to attention check items. *J. Res. Personal.* 82:103849

Dennis SA, Goodson BM, Pearson CA. 2020. Online worker fraud and evolving threats to the integrity of MTurk data: a discussion of virtual private servers and the limitations of IP-based screening procedures. *Behav. Res. Account.* 32(1):119–34

DeSimone JA, DeSimone AJ, Harms PD, Wood D. 2017. The differential impacts of two forms of insufficient effort responding. *Appl. Psychol.* 67(2):309–38

DeSimone JA, Harms PD. 2018. Dirty data: the effects of screening respondents who provide low-quality data in survey research. *J. Bus. Psychol.* 33(5):559–77

DeSimone JA, Harms PD, DeSimone AJ. 2015. Best practices and recommendations for data screening. *J. Organ. Behav.* 36(2):171–81

Dunn AM, Heggestad ED, Shanock LR, Theilgard N. 2018. Intra-individual response variability as an indicator of insufficient effort responding: comparison to other indicators and relationships with individual differences. *J. Bus. Psychol.* 33(1):105–21

Dupuis M, Meier E, Cuneo F. 2019. Detecting computer-generated random responding in questionnaire-based data: a comparison of seven indices. *Behav. Res. Methods* 51(5):2228–37

Etikan I, Musa SA, Alkassim RS. 2016. Comparison of convenience sampling and purposive sampling. *Am. J. Theor. Appl. Stat.* 5(1):1–4

Fleischer A, Mead AD, Huang J. 2015. Inattentive responding in MTurk and other online samples. *Ind. Organ. Psychol.* 8(2):196–202

Gibson AM, Bowling NA. 2020. The effects of questionnaire length and behavioral consequences of careless responding. *Eur. J. Psychol. Assess.* 36(2):410–20

Godinho A, Schell C, Cunningham JA. 2020. Out damn bot, out: recruiting real people into substance use studies on the internet. *Subst. Abuse* 41:3–5

Gordon ME, Slade LA, Schmitt N. 1986. The "science of the sophomore" revisited: from conjecture to empiricism. *Acad. Manag. Rev.* 11(1):191–207

Griffin M, Martino RJ, LoSchiavo C, Comer-Carruthers C, Krause KD, et al. 2021. Ensuring survey research data integrity in the era of internet bots. *Qual. Quant.* 56:2841–52

Hauser DJ, Schwarz N. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav. Res. Methods* 48(1):400–7

Highhouse S. 2009. Designing experiments that generalize. *Organ. Res. Methods* 12(3):554–66

Huang JL, Bowling NA, Liu M, Li Y. 2015a. Detecting insufficient effort responding with an infrequency scale: evaluating validity and participant reactions. *J. Bus. Psychol.* 30:299–311

Huang JL, Curran PG, Keeney J, Poposki EM, DeShon RP. 2012. Detecting and deterring insufficient effort responding to surveys. *J. Bus. Psychol.* 27(1):99–114

Huang JL, Liu M, Bowling NA. 2015b. Insufficient effort responding: examining an insidious confound in survey data. *J. Appl. Psychol.* 100(3):828–45

Johnson JA. 2005. Ascertaining the validity of individual protocols from web-based personality inventories. *J. Res. Personal.* 39:103–29

Kam CCS, Chan GHH. 2018. Examination of the validity of instructed response items in identifying careless respondents. *Personal. Individ. Differ.* 129:83–87

Kan IP, Drummey AB. 2018. Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce. *Comput. Hum. Behav.* 83:243–53

Keith MG, Tay L, Harms PD. 2017. Systems perspective of Amazon Mechanical Turk for organizational research: review and recommendations. *Front. Psychol.* 8:1359

Kennedy R, Clifford S, Burleigh T, Waggoner PD, Jewell R, Winter NJG. 2020. The shape of and solutions to the MTurk quality crisis. *Political Sci. Res. Methods* 8:614–29

Kim DS, McCabe CJ, Yamasaki BL, Louie KA, King KM. 2018. Detecting random responders with infrequency scales using an error-balancing threshold. *Behav. Res. Methods* 50(5):1960–70

Kung FY, Kwok N, Brown DJ. 2018. Are attention check questions a threat to scale validity? *Appl. Psychol.* 67(2):264–83

MacInnis CC, Boss HCD, Bourdage JS. 2020. More evidence of participant misrepresentation on MTurk and investigating who misrepresents. *Personal. Individ. Differ.* 152:109603

Mahalanobis PC. 1936. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* 12:49–55

Marcus B, Weigelt O, Hergert J, Gurt J, Gelléri P. 2017. The use of snowball sampling for multisource organizational research: some cause for concern. *Pers. Psychol.* 70(3):635–73

McGonagle AK, Huang JL, Walsh BM. 2016. Insufficient effort survey responding: an under-appreciated problem in work and organisational health psychology research. *Appl. Psychol.* 65(2):287–321

Meade AW, Craig SB. 2012. Identifying careless responses in survey data. *Psychol. Methods* 17(3):437–55

Moss AJ, Litman L. 2018. After the bot scare: understanding what's been happening with data collection on MTurk and how to stop it. *CloudResearch Blog*, Sept. 18. **https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/**

Moss AJ, Rosenzweig C, Jaffe SN, Gautam R, Robinson J, Litman L. 2021. Bots or inattentive humans? Identifying sources of low-quality data in online platforms. PsyArXiv wr8ds. **https://doi.org/10.31234/osf.io/wr8ds**

Newman A, Bavik YL, Mount M, Shao B. 2021. Data collection via online platforms: challenges and recommendations for future research. *Appl. Psychol.* 70(3):1380–402

Oppenheimer DM, Meyvis T, Davidenko N. 2009. Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* 45(4):867–72

Peer E, Vosgerau J, Acquisti A. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* 46(4):1023–31

Pozzar R, Hammer MJ, Underhill-Blazey M, Wright AA, Tulsky JA, et al. 2020. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. *J. Med. Internet Res.* 22:e23021

Ran S, Liu M, Marchiondo LA, Huang JL. 2015. Difference in response effort across sample types: perception or reality? *Ind. Organ. Psychol.* 8(2):202–8

Robinson J, Rosenzweig C, Moss AJ, Litman L. 2019. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLOS ONE* 14(12):e0226394

Sears DO. 1986. College sophomores in the laboratory: influences of a narrow data base on social psychology's view of human nature. *J. Personal. Soc. Psychol.* 51(3):515–30

Simone M. 2019. Bots started sabotaging my online research. I fought back. *First Opinion Blog*, Novemb. 21. **https://www.statnews.com/2019/11/21/bots-started-sabotaging-my-online-research-i-fought-back**

Storozuk A, Ashley M, Delage V, Maloney EA. 2020. Got bots? Practical recommendations to protect online survey data from bot attacks. *Quant. Methods Psychol.* 16(5):472–81

Tay L, Meade AW, Cao M. 2015. An overview and practical guide to IRT measurement equivalence analysis. *Organ. Res. Methods* 18(1):3–46

Teitcher JE, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. 2015. Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. *J. Law Med. Ethics* 43(1):116–33

Waggoner PD, Kennedy R, Clifford S. 2019. Detecting fraud in online surveys by tracing, scoring, and visualizing IP addresses. *J. Open Source Softw.* 4(37):1285

Wood D, Harms PD, Lowman GH, DeSimone JA. 2017. Response speed and response consistency as mutually validating indicators of data quality in online samples. *Soc. Psychol. Personal. Sci.* 8(4):454–64