An Inconvenient Truth: Arbitrary Distinctions Between Organizational, Mechanical Turk, and Other Convenience Samples

Richard N. Landers
Old Dominion University

Tara S. Behrend
The George Washington University

Sampling strategy has critical implications for the validity of a researcher’s conclusions. Despite this, sampling is frequently neglected in research methods textbooks, during the research design process, and in the reporting of our journals. The lack of guidance on this issue often leads reviewers and journal editors to rely on simple rules of thumb, myth, and tradition for judgments about sampling, which promotes the unnecessary and counterproductive characterization of sampling strategies as universally “good” or “bad.” Such oversimplification, especially by journal editors and reviewers, slows the progress of the social sciences by considering legitimate data sources to be categorically unacceptable. Instead, we argue that sampling is better understood in methodological terms of range restriction and omitted variables bias. This considered approach has far-reaching implications because in industrial–organizational (I-O) psychology, as in most social sciences, virtually all of the samples are convenience samples. Organizational samples are not gold standard research sources; instead, they are merely a specific type of convenience sample with their own positive and negative implications for validity. This fact does not condemn the science of I-O psychology but does highlight the need for more careful consideration of how and when a finding may generalize based on the particular mix of validity-related affordances provided by each sample source that might be used to investigate a particular research question. We call for researchers to explore such considerations cautiously and explicitly both in the publication and in the review of research.

Despite its importance, external validity receives only cursory treatment in most research methods textbooks. Even handbooks are light on coverage; in their lengthy, classic reference work on research methods, Pedhazur

Richard N. Landers, Department of Psychology, Old Dominion University; Tara S. Behrend, Department of Organizational Sciences, The George Washington University.

Correspondence concerning this article should be addressed to Richard N. Landers, 250 Mills Godwin Building, Department of Psychology, Old Dominion University, Norfolk, VA 23529, rnlanders@odu.edu
and Schmelkin (1991) devoted a scant four of their 819 pages to external validity, whereas Shadish, Cook, and Campbell (2002) provided a comparatively rigorous treatment with approximately 10 pages. In part because internal validity is a prerequisite for external validity, much greater focus in these volumes is placed on internal validity and the various techniques to maximize it. This bias is reflected more broadly in the writing and review process as well; although authors will devote pages of text to measure identification, experimental design, and analytic strategy, considerations related to external validity are often limited to a token paragraph in a limitations section. Given this focus in both seminal texts and current research literature, we suspect the balance of coverage in graduate education is similarly skewed.

With such a lack of guidance, many journal reviewers are left to rely on their own idiosyncratic reasoning as to whether external validity is threatened in a particular study. The use of obvious convenience sampling, in particular, is a lightning rod for criticism of study generalizability. Each time a new and more convenient sampling technique begins to gain traction, discussion begins anew regarding “appropriate” sampling. Arguments against sampling that is “a little too convenient” appear anew on scholarly discussion lists and in critical reviews returned by journal editors. At least one industrial–organizational (I-O) psychology journal, the Journal of Vocational Behavior, forbids the publication of research conducted with such samples, stating that the use of online panels “threaten[s] the integrity of research samples and the validity of results” (Elsevier, 2014, Introduction section, para. 4). Unfortunately, most of these arguments are based on neither empirical evidence nor a compelling theoretical model of validity or generalizability. Instead, they more typically rely on myth, intuition, and tradition. Comments like “This study is weakened by its reliance on a college student sample” are common. This kind of uncritical and nonspecific condemnation is harmful because simple decision rules categorizing particular sources of convenience samples as good or bad unnecessarily limit the types of samples from which researchers are willing to draw. Shrinking the pool of legitimate data sources this way slows scientific progress without cause. Researchers must consider threats to external validity systematically and scientifically.

As an opening salvo on external validity myths in this article, we explore an aspect of external validity that is a common to all research studies: generalizability from a sample to a desired population as driven by sampling strategy. The most traditional advice given in this domain is to carefully consider the population of interest, identify a sample that is a more or less random representation of that population, and then approach that sample with research participation requests. To the extent that response rates approach 100% with such a design, researchers generally conclude that external validity is not a concern. In practice, sampling is much messier. True random
sampling and even quasirandom sampling are sufficiently uncommon in the I-O psychology literature that we consider the researchers following this advice to be an exception to the common practice. More often, researchers conduct their studies with whatever sample is conveniently available, such as college students seeking extra credit, online panels seeking payment, or an organization for which the primary author happens to consult. We seek to provide some clarity as to the consequences of these practices.

When using such samples in the pursuit of empirically supporting a theory that is broadly applicable across organizations in I-O psychology, most researchers implicitly select “all employees in all organizations” as their population of interest. In fact, theories that only apply to specific types of organizations or jobs are often marginalized; few I-O psychology theories, for example, apply to “the psychology of retail sales workers.” A brief glance at recent issues of the Journal of Applied Psychology and Personnel Psychology reveals article titles and subtitles with broad themes like “Relationship of work motivations and behaviors to within-individual variation in the five-factor model of personality” and “Do financial rewards for performance enhance or undermine the satisfaction from emotional labor?” Such breadth of phrasing demonstrates that the intent of most I-O psychology researchers is to develop conclusions that can be applied broadly across jobs, industries, and cultures (hereafter referred to as the global worker pool), ostensibly and appropriately to maximize the value of those conclusions to I-O psychology practitioners. Thus, any currently or potentially employed person falls within the population of interest to most I-O psychologists.1 Within the global worker pool, the use of any convenient sample of current employees rather than a probability sample of the global pool introduces two technical but well-defined challenges: omitted variables and classic range restriction. Reliance on tradition is not required to interpret the effects of such convenience.

Our goal in this article is thus to explore how, and under which conditions, the convenience samples that are commonly available to I-O psychologists, such as the use of Internet panels, college students, and specific organizations, do or do not harm the external validity of research studies when researchers draw conclusions about the global worker pool. We explore this with the hope that journal editors and reviewers alike will suppress their tradition-based reactions to nonorganizational sampling techniques, such as the use of college students or Amazon Mechanical Turk (MTurk), and instead consider the specific strengths and weaknesses of all convenience

1 We note that there are exceptions to this statement, for example, the study of workers within high-stakes or extreme environments, atypical team structures, and so on and the fields of study that concern particular cultural or demographic characteristics. However, such research studies are much less common than are those drawing conclusions about overall worker behavior.
Historical Discussions of External Validity and Convenience Sampling

To understand the impact of convenience sampling on external validity, we must first carefully define each of these terms and explore their historical and current treatments, both explicit and implied.

**External Validity**

Major reference works vary widely in their definitions and interpretations of external validity. We cover the three works we believe to be the most common in I-O psychology graduate education: Pedhazur and Schmelkin (1991), Shadish et al. (2002), and Sackett and Larson (1990).

Pedhazur and Schmelkin (1991) defined external validity as “the generalizability of findings to or across [emphasis added] target populations, setting, times, and the like” (p. 229). In this consideration of validity, generalizing to is treated quite narrowly, and the term refers to the ability of a researcher to make valid conclusions about the population from which a particular sample is randomly drawn. In the case of convenience sampling, a researcher is generalizing to when drawing conclusions from the convenient sample to the convenient population from which it has been randomly or semirandomly drawn (e.g., from college students taking part in a study to a population of college students seeking extra credit during a particular semester). In contrast, generalizing across refers to a researcher’s ability to draw conclusions about a desirable population from a given nonprobability sample (e.g., from college students taking part in a study to human behavior in general). As previously described, generalizing to is rarely the external validity goal of I-O psychology researchers. Instead, convenient organizations are used to draw conclusions about organizations in general; convenient college students are used to draw conclusions about people in general. Thus, we are more specifically concerned with generalizing across from convenient samples to less convenient, but related populations.

Within this framework, Pedhazur and Schmelkin (1991) identified five major threats to external validity, one of which is specifically relevant to generalizing across, called the treatments–attributes interaction. This threat generally refers to any way in which the people studied are different from the population of interest. To remedy this problem, Pedhazur and Schmelkin (1991) stated that such attributes “must be included in the design” (p. 230). If such attributes cannot be included, researcher conclusions must be qualified with this limitation. Specific steps to accomplish this inclusion are not described.
Shadish et al. (2002) defined external validity as “the inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes” (p. 83). Using this definition, they next defined five broad objectives when generalizing from the results of a single study, the first of which is relevant to our purpose here: narrow to broad, which concerns drawing conclusions about populations from a research sample. When convenience sampling is criticized, it is typically for this reason; the research consumer believes that the “convenience” of the sample makes the sample a poor representation of the population.

Within this framework, external validity is threatened in one of five ways, two of which are relevant to the narrow to broad generalization objective. First, the causal relationship may interact with particular sample characteristics. Relevant to I-O psychology, Shadish and colleagues (2002) provided an example of this by describing an experiment in which the most highly qualified applicants to a work program were selected to demonstrate the positive effects of that program, enhancing the program’s apparent effectiveness. Second, there may be unknown context-dependent moderators. For example, in a study of the effectiveness of self-regulation interventions on web-based learning, students in a lab setting may generally be more willing to obey researcher instructions than an organization’s employees may be when those employees are spending time away from other, more pressing work while engaging in such training.

Sackett and Larson (1990) provided the most prominent advice tailored to I-O psychologists on this issue. Those authors defined external validity as concerning “the degree to which the results obtained in a given study would hold at others times, in other settings, or with other individuals” (Sackett & Larson, 1990, p. 430), which is primarily driven by Cook and Campbell’s (1976) definition. They spoke to the issue of convenience sampling specifically by stating, “because the participants and/or settings are not drawn at random from the intended target population and universe, respectively, the true representativeness of a convenience sample is always unknown. For this reason, representativeness cannot be used as a criterion for preferring one convenience sample over another” (pp. 422–433). Instead, they recommended two alternative criteria to explore the external validity of convenience samples: sample relevance and sample prototypicality. Sample relevance refers to the degree to which membership in the sample is defined similarly to membership in the population. They provided an example of irrelevance by describing a study of executive decision making conducted with a college student sample. With sample prototypicality, sample relevance is assumed. Sample prototypicality refers to the degree to which a particular research case is common within a larger research paradigm. They provided an example of prototypicality by describing a study exploring the predictive
validity of integrity tests; although a sample of senior executives completing such tests could be collected, a sample of nonmanagerial employees would be more prototypical of when such integrity tests would actually be employed. As such, they recommended triangulation; as researchers in the field come to understand more about the most prototypical groups, investigations should emphasize other, less-studied groups. This process of triangulation is essential to the accumulation of knowledge.

On the matter of external validity threats, Sackett and Larson (1990) stated that there are specific circumstances under which generalizability, in general, is of less concern. One of these circumstances is particularly relevant to the question of the generalizability of convenience samples to the global worker pool: “when a study is conducted solely for the purpose of testing a theory . . . [in a] falsificationist orientation” (Sackett & Larson, 1990, p. 435). Because this orientation is the current dominant approach within I-O psychology research, this argument is of particular note. Within the falsificationist orientation (i.e., research relying on null hypothesis significance testing), a theory is proposed and tested not to provide support for the theory but in an attempt to falsify it. The tested theory is treated as “given”; statistical tests that match the theory’s predictions “support” the theory, and statistical tests that do not match the theory’s predictions indicate that the theory is false. Theories thus remain until they are falsified or are changed in future empirical work. Within this framework, Sackett and Larson (1990) argued, “the sole criterion for selecting a setting and participant sample is that it be relevant—this is, that it fit within predefined population/universe boundaries” (p. 435).

Sackett and Larson (1990) also described two instances when external validity should not overshadow other characteristics of the study’s design, both of which relate to theory testing. If the primary question of interest is whether a phenomenon can occur, rather than whether it does occur, or how frequently it occurs, internal validity is of greater importance than is external validity. Similarly, if the purpose of a study is to falsify a theory, internal validity should take precedence; in such cases, reasonable sacrifices to external validity are justifiable.

From these three perspectives, we have identified a common core in regard to convenience samples and their effects on external validity. When sampling at random from a population, researchers are able to rely on classical test theory to support the argument that a sample is reasonably representative of a target population. When convenience sampling, which is the approach of virtually all I-O psychology research, researchers cannot rely on probability alone when making the argument of representativeness. Instead, convenience sampling involves randomly sampling a convenient population and making a rational argument as to why the convenient population is
sufficiently similar to the intended population such that statistical conclusions about the convenient population may be used to draw theoretical conclusions about the intended population.

In our experience, such justification is rare in the I-O psychology literature. When college students are sampled, an argument should be made that the empirical relationships observed are likely to be similar in the global worker pool. Instead, a few lines in the limitations section (e.g., “These results may not generalize due to the use of a college student sample.”) are usually the extent of this argument. When a particular organization is chosen as the target, similar challenges are faced. There is no guarantee that the chosen organization’s employees, who have likely been subjected to selection procedures, training, onboarding, team building, and an eclectic mix of other interventions, represent the global worker pool. Yet the pros and cons of such a sampling choice are rarely, if ever, mentioned. There has been a recent push for organizational sciences journals to require authors to describe the context of their research, which will allow readers to judge external validity for themselves (e.g., Johns, 2006). Despite these calls, it is not typical to describe the context or sample in detail, except in the few journals explicitly requiring this (Rynes, 2012).

**Convenience Sampling**

Convenience sampling is often treated quite casually, even in organizational research methods textbooks. For example, Trochim and Donnelly (2008) devoted only a few paragraphs to the use of convenience sampling in their research methods textbook, mostly to note that such samples are both common and problematic. In reference to a hypothetical convenience sample, Trochim and Donnelly (2008) stated that “such a sampling plan almost guarantees that the sample selected will not represent the population” (p. 51) and elaborated that although convenience sampling may have some value for exploratory research, it has “perhaps the least usefulness for generalizability of findings” (p. 51). The authors drew from past political polls to give examples of incorrect election predictions that were based on the oversampling of wealthy, phone-owning voters, who tended to be disproportionately Republican. We concur with the authors’ emphasis on omitted variables that may moderate or limit findings. However, we contend that this example does not condemn convenience samples; rather, it serves as a reminder to researchers to investigate exactly which variables may be peculiar in a given sample. In the example referenced by these authors, a phone poll necessarily omitted voters who did not own a phone. Careful thought ahead of time would have brought this to light. Further, the question of interest in a political poll is “How often does this phenomenon occur?” (i.e., What is the prevalence of
it is not an appropriate use of convenience sampling, but this does not reduce the value of convenience samples when investigating other types of questions.

**Convenience Sampling in Modern I-O Psychology**

Issues of sample and research design are critical to the trustworthiness of research findings. Despite this, design issues are frequently treated with generic rules of thumb as opposed to considered argument. For instance, Aguinis and Vandenberg (2014) reported that in reviewer comments regarding articles published in *Organizational Research Methods*, only 15% of the comments addressed research design issues, whereas data analysis was addressed by half. Similarly, reviews for *Psychological Methods* addressed design only 10% of the time, but statistical issues were mentioned in 70% of reviews. Our own review of organizational research methods textbooks found that sampling was typically mentioned only briefly, without much consideration of the ways that sample characteristics and recruitment methods may influence validity.

Convenience samples are not selected at random. Their external validity depends on the particular characteristics of the sample and the setting and procedures of the research. All samples in I-O psychology have peculiar characteristics. This should not condemn these samples. Instead, a serious consideration of how these characteristics moderate or limit the study findings should be conducted. For example, Staw and Ross (1980) found that managerial, business, and psychology students were sequentially less positive in the ratings of a hypothetical manager. One could interpret this pattern of findings as suggesting that students are not appropriate participants for studies involving manager ratings. However, Staw and Ross (1980) considered that the distinguishing variable between samples was experience with the manager role. Students, who had the least experience with management, behaved differently from managers, who had the most experience. Instead of disregarding evidence from student samples, this interpretation illuminated a key variable of interest and progressed knowledge in this area. This study also serves as an example of the ways that sample characteristics are distinct from lab versus field choices.

The lab versus field controversy is the issue related to external validity that is most extensively discussed by I-O psychologists. Dipboye (1990) noted that the field’s acceptance of laboratory research ebbs and flows over time, citing patterns observed in the *Journal of Applied Psychology* in the 1970s and 1980s. He reviewed comparisons between lab and field with regard to validity, especially generalizability, which tends to be the main complaint about lab research. It is interesting to note that many of the primary complaints lodged about lab research tended to describe “student” and
nonstudent samples (see Campbell, 1986; Gordon, Slade, & Schmitt, 1986; Ilgen, 1986), conflating this distinction with lab versus field. It is possible that in the 1980s when this controversy was at its most heated, this was an accurate representation. However, this assumption would be unwise in the present day.

Dipboye (1990) also noted that field research at that time was extremely limited, with an oversampling of professional/managerial employees and military personnel and an undersampling of clerical and blue-collar professions. Field research at this time was also more likely to rely on self-report measures than on behavioral observations. Both of these patterns limit the generalizability of findings from field research.

Dipboye (1990), agreeing with McGrath and others (McGrath, 1982; McGrath & Brinberg, 1984) who have written about this topic, noted that “the only hope for building a valid body of knowledge on organizational behavior is to use a diversity of settings and strategies... the ideal mix depends on the phenomenon under investigation, the skills of the investigators, and the availability of research settings” (p. 12). Although not explicitly stated, this can be interpreted as an endorsement of varied convenience samples. We hope to give the same scrutiny to various types of convenience samples that researchers before us have given to lab versus field settings.

We suspect that one driver of this debate has little to do with validity. Rather, there is a perception from practitioners that academics do not focus on questions that are important in the day-to-day functioning of organizations (see, e.g., Rynes, 2012; Rynes, Bartunek, & Daft, 2001). An academic researcher who conducts lab studies may be perceived as being out of touch with organizational reality. This may be a fair criticism of many ivory tower researchers, but such evaluations should not be based on sampling decisions alone.

Sackett and Larson (1990) gave specific attention to the choices researchers must make with regard to who will participate in a research study, noting that not all choices made are conscious ones. The availability of resources often guides the selection of participants. This is the very definition of a convenience sample; it is convenient. This means that an organization with which a researcher has a relationship is more convenient than one that is resistant. In addition, norms and fads in a field may influence the choices a researcher makes (Rynes, 2013), and some choices may be met with more scrutiny when they do not conform with these fads (Sackett & Larson, 1990). As such, various forms of the generalizability conversation have taken place within I-O psychology over the past decades, and some are just emerging now. We describe each major type of convenience sample common to I-O psychology below, including college student, online panel, crowdsourced, organizational, and snowball samples.
College Students
Many people equate the terms *convenience sample* and *college student*, whether the sample is drawn from a psychology participant pool or a class of business students. The concern about the external validity of these samples has been so great that there is a sizable literature comparing students and nonstudents in areas such as personnel selection and performance appraisal (Gordon et al., 1986; Greenberg, 1987; Landy, 2008). Landy (2008) went so far as to state student samples are ruining the credibility of research relating to stereotypes and bias in employment decisions because of this area’s heavy reliance on students who presumably behave differently from nonstudents with regard to stereotypes and decision making.

Concerns about college student samples are quite common in I-O psychology literature. For example, Aguinis and Edwards (2013) stated that “designs that allow for researcher control, random assignment, and manipulation of variables yield high levels of confidence regarding internal validity but are usually weaker regarding external validity, e.g., due to the use of sophomores in university laboratory settings [emphasis added]” (p. 158). The authors did equivocate somewhat, using “usually” and “presumably” to describe the ways that student samples affect generalizability while also invoking the memorable but misleading phrase typically used to criticize fields relying on such samples: “science of the sophomore.”

In our experience, master of business administration student samples receive less scrutiny than do psychology student samples. It is probably the case that researchers assume that the average master of business administration student has more work experience than does the average psychology undergraduate, although this is rarely discussed in articles utilizing master of business administration students. Such reasoning would be relevant only in cases in which work experience in a corporate organization is relevant to hypotheses. The use of student samples consisting primarily of older, nontraditional college students would likely compensate for any relative limitations in samples of traditional college students.

Online Panels
Online panels, frequently used by market researchers, describe groups of people who volunteer or groups of people who are paid to be available to complete questionnaires. Participants often provide demographic or other information to panel organizers who make this information available to researchers who wish to recruit participants fitting a particular criterion. Panel participants may opt in or out of any particular study, and they may make this decision on the basis of their interest in the research topic, their availability, or the compensation offered. Participants are usually paid a small fee to complete a research study, and researchers typically pay both the panel...
host and the participant. Online survey software companies Qualtrics and SurveyMonkey both offer panel and recruitment services to researchers as part of a more comprehensive project management package, which includes survey design and analysis.

StudyResponse is another popular panel for organizational researchers. Its creators describe its appropriateness as follows:

StudyResponse was created in response to difficulties that we and our colleagues had while trying to recruit participants for various types of studies that are not sensitive to so called “non-sampling errors” [emphasis added]. In particular, StudyResponse is likely to be useful for studies of phenomena that exhibit robust correlational relationships [emphasis added]. StudyResponse has also been used for qualitative data collections, where data would not be analyzed with statistical techniques. Generally speaking, StudyResponse is inappropriate for demography [emphasis added] and other applications where coverage errors could adversely bias your results. (StudyResponse.net, 2015)

This description is consistent with what we note above, namely, convenience samples in general should be viewed with the most scrutiny when a research question deals with estimating precise effect magnitudes or with measuring the prevalence of a phenomenon, as opposed to the possibility of a phenomenon existing. A number of published journal articles using StudyResponse data are listed on the StudyResponse web site, including one published in the Academy of Management Journal and one in the Journal of Personality Assessment (full list here: http://www.studyresponse.net/techreports.htm; Stanton & Weiss, 2002).

**Crowdsourcing/Amazon Mechanical Turk (MTurk)**

MTurk is a service offered by Amazon.com, Inc., that purports to “automate” difficult-to-automate tasks by splitting the work among many human workers. Amazon originally built the service for internal purposes, such as tagging product images. In this case, workers would view a picture of an Amazon product and generate descriptors (e.g., “lamp,” “blue,” “modern”). Workers would be paid a few cents per task. MTurk has since evolved to be open to requestors (those requesting work) and workers (those completing work) from any organization or location. Since at least 2009, social science researchers have used MTurk to recruit participants for a variety of topics and research designs (Behrend, Sharek, Meade, & Wiebe, 2011; Buhrmester, Kwang, & Gosling, 2011).

We believe this type of sample has great potential for organizational researchers. Aguinis and colleagues (Aguinis & Edwards, 2013; Aguinis & Lawal, 2012) referred to this type of service as “eLancing” and suggested that it is the ideal blend of an experimental control and a naturalistic setting. In fact, the use of MTurk may solve a problem that has vexed the research community for decades, namely, the severe oversampling of participants from Western, educated, industrialized, rich, and democratic (WEIRD)
backgrounds (Henrich, Heine, & Norenzayan, 2010). Given that a large proportion of MTurk users are from Asian countries, this service can provide samples of non-Western, nonrich participants with ease.

At present, the most substantial barrier to greater adoption of MTurk as a potential method for convenience sampling is reviewer unfamiliarity with and subsequent dismissal of the approach because of a variety of untested assumptions. In our experience, such reviewers do raise concerns worth considering, but the reviewers assume these issues are unique to MTurk as a data source or overestimate the impact of the issue on data quality. Specifically, these concerns can be grouped into four major themes, which we describe alongside sample comments from actual reviewers. We present these comments verbatim and anonymously to illustrate actual reviewer thinking regarding these issues. First, we have noted repeated participation concerns (e.g., “we are concerned that the median MTurk participant has completed forms for 30 different studies”). This is problematic only if repeated participation may harm validity; for example, we would not expect personality measures to become less accurate over repeated administrations. Second, we have noted concerns over compensation and resulting motivation (e.g., “The participants were paid $2 [which is a very small amount] to participate . . . one has to wonder since the primary motivation of individuals who volunteer is to earn money”). This is of concern only if the compensation level or financially driven motivation can be theoretically linked to effect size. Third, we have noted concern over selection bias (e.g., “What about participants who viewed the task but didn’t complete it? I find the lack of control over subjects troublesome”), but such issues are common in all convenience samples. Fourth, we have noted concerns over the relevance of the sample to working populations (e.g., “Perhaps MTurk could get you a more diverse sample than the typical student population, but at what cost?”), but for researchers with the goal of generalizing to the global worker pool, MTurk may be ideal for the reasons we outlined earlier. Consideration of the particular strengths and weaknesses of any convenience sampling approach is certainly warranted, but the weaknesses of a particular approach may or may not apply to a particular research question. As we note above, reliance on rules of thumb based on these or any other criteria is unwise.

Snowball and Network Samples

Snowball and network samples include e-mails distributed through social networks, alumni associations, civic groups, and referrals. Such samples may involve participants who are personal contacts of the researcher. This type of sample seems to be more common in qualitative research, which is usually less concerned with issues of generalizability. For instance, a researcher who wishes to generate a list of possible job search motives may do so by
interviewing people who have signed up with a university career services office. In this case, the network characteristics are relevant and valuable to the study. It may be tempting for a researcher to use this type of sample for quantitative research. This is appropriate in only a limited number of cases: first, if the research question is concerned with the dynamics of the network itself (e.g., the job search example above, or understanding how communication occurs in social and professional networks); second, if the research question involves a highly specific and unusual subject matter expertise or characteristic that requires the use of personal contacts for recruitment (e.g., female airline industry executives); or third, if the behavior of interest is both infrequent and hard to track, but participants have knowledge of others who fit the criteria for the study (e.g., undisclosed workplace relationships).

**Organizational Samples**

The preceding sample types are occasionally met with skepticism by organizational researchers, who believe that organizational samples are more generalizable than are nonorganizational samples. Indeed, this is the most typical convenience sample reported in I-O psychology journals. Because this point may seem surprising, it bears repeating: The most typical convenience sample found in I-O psychology journals involves a single organization with which the researcher has some prior relationship. With only a few notable exceptions, in no published I-O psychology research that we could locate had researchers solicited employees drawn at random from the full theoretical population of employees across all organizations of interest. Thus, nearly all I-O psychology research currently published is based on convenience samples, and that research should be explicitly considered as such. In most cases, any idiosyncratic characteristics of these organizations, which may or may not bias results, are left unreported and unknown.

**Statistical Effects of Convenience Sampling**

Sampling strategies can affect the conclusions one draws from convenient research samples in two major ways. First, samples may be range restricted on a variable or variables of interest; for example, a sample consisting of engineers may be range restricted in cognitive ability. Second, a sample may have characteristics that either covary or interact with the variables in a model; to the extent that these variables are not accounted for, biasing effects may occur. Below, we describe these two concepts and consider how they may manifest in the types of convenience samples common to I-O psychology when researchers are attempting to draw conclusions about the global worker pool.

---

2 One such exception is Wang and colleagues’ work (e.g., Wang & Hanges, 2011; Wang & Russell, 2005; Wang, Zhan, Liu, & Shultz, 2008). In these studies, the researchers used stratified random samples drawn from U.S. census data.
Range Restriction

In I-O psychology, range restriction is most commonly discussed in the context of psychometric meta-analysis (Hunter & Schmidt, 2004), especially related to the validity generalization debate (Schmidt et al., 1993). The basic concept of range restriction, however, is limited neither to selection nor to meta-analysis. Range restriction occurs whenever a sample variable’s range is reduced from its range in the population. When considering the relationship between two particular variables, this takes one of two general forms (Sackett & Yang, 2000). First, direct range restriction describes situations in which a hard cutoff has been imposed directly on a variable of interest. For example, consider an organization where a minimum cutoff score on muscular strength has been set at, say, the ability to lift a 40-pound weight while standing. If applicants do not meet this minimum standard, they will not be hired. Thus, current employees of this organization are directly range restricted on muscular strength. Second, indirect range restriction describes situations in which a cutoff has been imposed on another variable (measured or unmeasured) that is correlated with a variable of interest. For example, most American organizations incorporate an interview as part of their employee selection process, and in many cases, interview scores correlate with cognitive ability. Within a particular organization where such an interview was used, any researcher interested in drawing conclusions about cognitive ability would need to consider the effects of indirect range restriction. More broadly, nationality and culture can also be interpreted as range restriction; when citizens of a single nation are selected for inclusion in a study and others are excluded, any variable correlated with nationality will be biased in its generalization to the global worker pool because of indirect range restriction. Researchers typically ignore this limitation when reporting their results yet claim to have tested theories that generalize to the global pool.

Given this, range restriction in relation to the global worker pool occurs in nearly every study conducted within I-O psychology. With such a massive amount of range restriction going on, one might wonder what effect this has had on the external validity of I-O psychology research literature. Fortunately, the bias introduced by range restriction is well understood, and the bias can even be calculated mathematically under certain circumstances. In general, range restriction leads to attenuation of observed effect sizes, and direct range restriction leads to greater attenuation than does indirect range restriction. This is due to the more proximal nature of direct range restriction; when range restriction occurs further from a focal variable in its nomological net, the effects are buffered through one or more mediators. Calculation of the precise statistical effects and appropriate corrections is straightforward if sufficient contextual information is known about both
unrestricted and restricted samples, even if range restriction is indirect (Sackett, Lievens, Berry, & Landers, 2007).

**Range Restriction in Convenience Samples**

Each of the convenience samples described above will exhibit some degree of range restriction. College students are directly restricted in quantity of education and in whatever selection system is used by their college but also indirectly restricted in work experience, age, earning potential, cognitive ability, personality, and geographic location. Both online panels and crowdsourced work marketplaces like MTurk are directly restricted by the participant’s decision to join the website and indirectly restricted by anything correlated with that decision, such as earning potential, current employment, and motivation. Organizational job incumbents are directly restricted by the selection procedures in place at their organizations and indirectly restricted by anything correlated with those procedures or with attrition from that organization, such as job-related knowledge, job-related skills, cognitive abilities, physical abilities, personality, interests, and geographic location. All convenience sample sources, organizations included, potentially introduce range restriction. It cannot be avoided. Given that, the question that researchers must ask when choosing a sample is a specific one: Are the characteristics that are range restricted in the convenience samples we have access to likely to be correlated with the variables we are interested in measuring? If not, external validity will not be threatened for this reason.

**Omitted Variables**

Perhaps a more insidious, and certainly a more difficult to address, problem is that of omitted variables. A model can never contain all possible variables of interest; the purpose of a model is to produce maximum explanatory power with as few constructs and relationships as possible (i.e., parsimony). Including the entire universe of variables in a model does not serve to simplify anything. However, omitting variables introduces the possibility that the model is inaccurate; specifically, observed effects between predictors and criteria may be inflated or deflated because variance associated with the omitted variable is not considered. This problem has been described numerous times in the literature and has been referred to as omitted variables bias, or left out variables error (James, 1980; Mauro, 1990; Meade, Behrend, & Lance, 2009); it has also been described as one particular type of model misspecification in structural equation modeling (e.g., Kenny, 1979). The omitted variable may be either an additional predictor or a moderator of the observed effects in the model.

In the first case, the omitted variable is a predictor. This is also known as the third variable problem, wherein the omitted variable is a common
cause of both the predictor and the outcome in the model. In this case, the effect of the predictor will be overestimated to the extent that the predictor and omitted variable are related or will be underestimated in the event that the omitted variable is related to the predictor but not the outcome. The cause for concern is highest when the omitted variable relates strongly to the outcome and moderately with the predictors in the model. Meade and colleagues (2009) demonstrated that if the omitted variable is uncorrelated with the predictor, however, no bias occurs, either positive or negative.

If the unmeasured variable is a moderator, however, more concern is warranted. Observed effects may be reversed, nullified, or inflated, depending on the particular value of or any range restriction in the moderator. Sensitivity analysis can be conducted to determine the potential magnitude of unmeasured interaction effects and the robustness of the model. This is also an area in which meta-analysis can be useful in determining the likelihood of unmeasured moderators, pointing to the importance of a thorough description of the study context and sample in all published work (Johns, 2006).

**Omitted Variables and Convenience Samples**

An often unstated assumption with regard to convenience samples is that some characteristic of the sample is relevant to the model as a moderator or predictor and should thus be included in analyses. When reviewers raise concerns about convenience samples, they are often implicitly suggesting that an omitted variable is biasing the results. With regard to college students, the unmeasured variable may be work experience or consequences for poor performance on an experimental task. With regard to online samples, computer expertise may influence observed effects.

Four remedies to the omitted variables problem have been suggested, but only some of these remedies are relevant to sampling concerns. The first remedy is to use experimental control/random assignment, and the second is to model additional variables; these two suggestions do not address sampling concerns because presumably the sample characteristic is common to everyone in the sample, and it would not be possible to model the characteristic (e.g., including “college student status” in a college student sample would have no effect). Further, it is not advisable to model many sample characteristic variables without cause; we agree with Spector and Brannick (2010) on the misuse and overuse of control variables. The third and fourth remedies are potentially useful to consider. The third remedy is to use previous research to justify one’s assumptions; this remedy is important because it requires that there be a theoretical reason for why sample characteristics would be potentially problematic or not problematic. This is also the only available option if the omitted variable in question is a potential moderator. The fourth remedy is a careful consideration of the purpose of the research.
Meade and colleagues (2009) noted that left out variables error is more of a concern if one’s goal is to estimate precise path magnitudes and less of a concern if significance or effect sizes are the goal, even if the omitted variable in question has a moderate to large correlation with the predictor. This recommendation is consistent with Sackett and Larson’s (1990) advice regarding the appropriate type of research questions for convenience samples. Thus we recommend that careful consideration be given to the underlying theory as it relates to the variables in one’s model. For example, the use of a college student sample is a justifiably poor design decision when some characteristic of college students would be expected to relate to both the predictor and the outcome under study. Similar questions should be raised when considering single-organization samples; for example, researchers conducting a study that examines leader behaviors and follower outcomes in an organization with a positive organizational culture may wish to explore how organizational culture can drive both follower outcomes and leader behaviors. At a minimum, theoretically relevant characteristics of the organization should be described in sufficient detail for meta-analytic explorations to identify these sample characteristics.

**Recommendations and Considerations**

In this article, we present a historical treatment of external validity, present an exploration of convenience sampling as it is currently practiced in I-O psychology, and provide an overview of the statistical and interpretive implications of convenience. From this review, we have developed five specific recommendations for the use and critique of convenience sampling in I-O psychology.

**Recommendation 1. Don’t Assume Organizational Sampling Is a Gold Standard Sampling Strategy**

Because I-O psychology focuses on the description, explanation, and prediction of worker behavior, it is quite natural to assume that the “best” samples would come from organizations. However, this assumption comes from an uncritical consideration of precisely what organizational samples have to offer. Organizational samples are not probability samples and should not be treated as such. In fact, organizational samples are often quite limited in ambiguous and difficult-to-measure ways. Not only are employees within an organization range restricted on whatever selection devices were used to hire them, but there are also a host of omitted variables at higher levels of analysis that may influence the results of a particular study (e.g., organizational culture, industry, country).

Online convenience samples in fact provide a number of advantages over traditional convenience sampling. For example, researchers have criticized traditional college student and organizational convenience samples as being
WEIRD (Henrich et al., 2010), limiting their applicability outside of WEIRD populations. This is a problem readily solved with the use of participants drawn from sources like MTurk, which have a sizable international membership. If we intend to create theory broadly applicable across organizational contexts, MTurk and similar samples may prove superior to those collected from single convenient organizations. Because most researchers are WEIRD and consult for WEIRD organizations, most of their research is WEIRD too.

As an example of problems potentially faced in organizational samples, consider the following study of a leadership training intervention. Our goal in this study is to conclude, “Does this intervention help leaders perform more effectively?” In this quasi-experimental study, two divisions of an organization are randomly assigned to conditions: Division A is assigned to receive the training intervention and Division B is assigned to act as a control. Divisions must be assigned instead of individuals because there is no way to isolate leaders within a division from one another; the validity of the study manipulation would be threatened. Because of this organizational reality, the internal validity of the study has been weakened. Furthermore, because Division A leaders interact with each other a great deal, they take this opportunity to compare notes and improve their application of the training. This adds additional confounds to the design. If the intervention was provided to a single individual, it may not have worked as well, if at all. If the intervention was used in an organization with weaker leaders, it may not have worked as well, if at all.

Now consider, in contrast, if this study had been conducted using an online panel, asking individuals in supervisory roles to try out a leadership intervention. Some aspects of the training experience are lost because it must now be delivered online—a distinct downside—but the diversity of participants in this online panel means that it is unlikely that more than one leader will try out these leadership skills within a particular organization. This avoids some of the internal validity problems faced when using the organizational sample. By using the online panel as a screening survey, researchers can collect data from a broad cross-organizational sample without the omitted variables problems faced within an organization. However, it will also be substantially more difficult to collect surveys from direct reports.

Neither approach is obviously preferable, and that challenge is at the heart of this recommendation. In this scenario, the researcher would need to consider the trade-offs between the two approaches. The organization is not automatically superior. Is the use of a more authentic leadership training session (in-person versus online) and the increased ease of collecting data from direct reports worth the sacrifice in both internal and external validity? This is a decision the researcher must make in either case and defend in his or her article.
Recommendation 2. Don’t Automatically Condemn College Students, Online Panels, or Crowdsourced Samples

One of the primary conclusions here, the one that we most hope researchers will adopt, is that convenience sample is not synonymous with poor sampling strategy. Virtually all samples used in I-O psychology are convenience samples. Instead, we urge researchers to identify any range restriction likely introduced by the sampling strategy of that convenience sample, to determine whether any moderators of study effects have been omitted, to explore the potential impact of these issues given the research questions, and to choose which convenience sample will meet the research goals.

A particularly notable advantage of online panels and crowdsourced samples is that they are not necessarily WEIRD. Broad international samples can be collected as a basic feature of these approaches. Even when such sampling strategies are restricted to participants from a particular country, the participants may be more representative of that country’s worker pool than workers in one particular organization would be.

Another advantage of MTurk over other convenience samples becomes more apparent when one considers the nature of range restriction in each sample. In MTurk samples, there is only one convenient population to be concerned about. Researchers can map out the nature of range restriction in MTurk samples drawn from that population (i.e., which variables are restricted in comparison with the global worker pool) and build an empirical literature describing those differences. Each study conducted on MTurk adds knowledge about such parameters. In contrast, in individual organizations, the responsibility for this exploration lies entirely with the researcher sampling from that organization. Ideally, any researcher conducting research within a single organization would review global norms for all of the variable measures of interest and compare these variables with the range of those variables within the sample, reporting this information in any submitted articles that are based on this sample. This places a much greater burden on researchers, reducing the value of organizational samples when researchers are aiming to ensure high external validity.

With regard to both online panels and crowdsourced marketplaces such as MTurk, there are several instances in which this type of sample is not only acceptable—it is also ideal. Reis and Gosling (2010) outlined a number of such instances in the field of social psychology, such as the study of online behavior. In the field of I-O psychology, there are many phenomena that take place exclusively on the Internet. As such, the Internet is the best place to recruit and study individuals who engage in those phenomena. As one example, it may be possible to determine what makes a recruitment ad effective, and for whom, by manipulating the language in an MTurk advertisement and tracking signup rates.
**Recommendation 3. Don't Assume That “Difficult To Collect" Is Synonymous With "Good Data”**

There is a natural tendency to consider the difficulty of collecting a sample and to consider that difficulty to be a proxy of sample quality. For example, we have seen MTurk criticized as being “a little too convenient.” It is important to remember that convenience is not a single continuum on which convenient is bad and inconvenient is good. Instead, each sampling strategy brings its own advantages and disadvantages along multiple dimensions of value in terms of both internal and external validity. These must be considered explicitly.

**Recommendation 4. Do Consider the Specific Merits and Drawbacks of Each Convenience Sampling Approach**

As highlighted throughout this article, simple rules of thumb should not be used to praise or condemn any particular convenient source of data. Instead, we recommend a five-step process:

1. Explore prior theory and identify related constructs in the broader nomological net of all constructs being studied.
2. Identify any variables in the target convenience sample likely to be range restricted or to have an atypical mean both at the level of the study (e.g., individual, team) and above it (e.g., team, organization, industry, nation). In organizational samples, researchers should consider the existing selection systems, the organizational culture (including leadership), and the industry/work domain in particular. In online and college student samples, selection and motivational variables are of concern.
3. Decide whether prior theory suggests any interactions between any variable within the nomological net of the study’s constructs and the characteristics of the sample.
4. Consider all potential trade-offs as a result of these interactions and choose the sample that best addresses stated research questions. Unless there is probability sampling, there will always be trade-offs.
5. Describe all of this reasoning in any submitted article.

We also recommend that editors do not request the removal of such reasoning in an effort to save printing space.

**Recommendation 5. Do Incorporate Recommended Data Integrity Practices Regardless of Sampling Strategy**

A number of best practices have been generated over time to increase the likelihood of obtaining high quality data regardless of sampling strategy. Meade and Craig (2012) provided one of the most comprehensive explorations of these approaches currently available, recommending the use of
instructed response items (e.g., questions stating “Please response ‘Strongly Disagree’ to this question”), consistency indices, and outlier analyses. Such practices are essential because online convenience samples in particular are criticized because of reviewer perceptions that these samples provide lower quality data. However, the relative base rates of careless responders among organizational, college student, and online samples is an empirical question, and it can only be resolved by conducting more research tracking the indicators like those recommended by Meade and Craig, in such samples.

Conclusion
In closing, we highlight these issues not to condemn organizational samples or, more broadly, convenience samples but instead to call on I-O psychology researchers to more carefully consider the nature of convenience in this new age of big data and creative Internet sampling, especially as drawn from MTurk. Simple decision rules categorizing particular sources of convenience samples as good or bad ultimately harms researchers’ ability to conduct good science by limiting the types of samples from which researchers are willing to draw. Scaring researchers away from these sources slows scientific progress unnecessarily. Sample sources like MTurk and other Internet sources are neither better nor worse than other more common convenience samples; they are merely different. In all cases, we should consider these differences explicitly and scientifically.

References

3 See Behrend et al. (2011) for an example of how these comparisons could be conducted.


