1 Student Samples in Research

Michael Basil

Abstract

This chapter provides an overview on the use and validity of student samples in the behavioral and social sciences. In some instances, data collected from students can be of limited value or even inappropriate; however, in other cases, this approach provides useful data. I offer three general ways to evaluate the use of student samples. First, consider the research design. Descriptive studies that rely on students to draw inferences about the overall population are likely problematic. Second, statistical controls such as multivariate analyses that adjust for other factors may reduce some of the biases that may be introduced through sampling. Third, consider the theorized mechanism – a clear theoretical mechanism that does not vary based on the demographics of the sample allows us to put more faith in constrained samples. Despite these approaches, and regardless of our methods, statistics, and theoretical mechanism, we should be cautious with generalizability claims.

Keywords: Sampling Students, Generalizability, Validity, Surveys, Experiments

Although some research relies on a census of an entire population, for the last 200 years there has been an increased reliance on examining a limited portion of a population (i.e., a sample) to try to understand the overall population (Fienberg & Tanur, 1996; Kruskal & Mosteller, 1980). This concept is not completely novel; as an example, Stephan (1948) points to the long history of the simple act of tasting or testing a small portion of a liquid. Over these past 200 years, the use of sampling a portion of a population has been applied in a wide variety of fields and in a wide variety of situations by researchers throughout the social and behavioral sciences. The primary reasons scientists use samples are the convenience, accessibility, and lower costs of research (Espinosa & Ortinau, 2016).

This chapter will focus on the use of student samples in research, specifically the validity of sampling college or university students. As mentioned earlier, reliance on samples to make inferences about the overall population can be traced to efforts in the late nineteenth and early twentieth century, involving Nicolai Kiaer, Jerzy Neyman, and R. A. Fisher (Fienberg & Tanur, 1996; Kruskal & Mosteller, 1980). The notion of sampling was likely originally conceptualized to examine the effects of agricultural experiments, but as the social and behavioral sciences evolved, the use of statistics has been applied to allow researchers to take the results from samples of

people and draw inferences about the general population. The field of inferential statistics developed so that we can derive probability-based confidence intervals for the likelihood of the results holding for the overall population. Statistically speaking, however, confidence intervals are predicated on the notion that the population was sampled at random, with no systematic biases (Kish, 1957; Tabachnick & Fidell, 2018). When samples are not simple random samples, these confidence intervals are often biased.

In this chapter, I propose three critical questions to consider about the use of student samples and some broad guidelines on interpreting the validity of data collected from homogeneous samples in general, with a special focus on college or university students. To evaluate the use of student samples, three questions should be posed. First, what is the research method? Descriptive surveys must be representative of the overall population they wish to represent. If the sample is composed entirely of students, then extrapolating to other student groups may or may not be reasonable, depending on the student body we are talking about, and drawing inferences to the general public may be even more problematic. For example, results of a study on sexual attitudes and behaviors based on a sample from a small religious college consisting of younger unmarried people likely will not bear much resemblance to a commuter college with an older, married population, and drawing inferences about the general public is likely even more problematic. In the case of most experiments, because a manipulation is randomly assigned to a group and the effect is measured, representativeness is often believed to be less important (Falk et al., 2013). Specifically, experimental research demonstrates whether the manipulation affects the sample. Inferential statistics provide confidence intervals showing the likelihood that the observed difference would hold to the population from which the sample was drawn (Edgington, 1966). Although we cannot say definitively whether the same effect would hold for a different population, we can largely rule out self-selection bias and third variables as possible sources of bias.

A second question to ask when using student samples is whether there are any statistical controls over other factors. Simple descriptive analyses are most susceptible to bias through sampling. Increasingly, however, due to common practice and because of increased computing power, more research relies on bivariate or multivariate statistical analyses with multiple independent variables on the outcome measures. This reduces some of the potential biases involved or allows them to be detected. For example, including students' religious orientation, age, and whether they are married would likely reveal insights into the differences in students' sexual attitudes and behaviors between religious and commuter colleges. If there is a difference in sexual attitudes that can be explained by religion, age, or marital status, this can be discovered through statistical analyses and would help when interpreting the findings. This approach is not as powerful as random assignment but does provide a means of assessing whether other factors may be biasing our findings (see Chapter 10 in this volume).

The third question to ask about sampling is whether the nature of the sample itself is likely to affect the theoretical mechanism in some identifiable way. When studying a particular phenomenon, is there reason to believe that the underlying process or

mechanism behind the phenomenon would be different for different groups of the population? If there is, then we should probably not be using a student sample, or at least should not seek to generalize the results beyond the sampled population. To generalize such a study, it would be necessary to replicate the findings with a different sample. This question is largely a matter of logic and theory but can be guided by previous research in that area. Further, making specific a priori predictions about the underlying process and measuring that helps us better understand the process and reduce the likelihood of Type I errors.

Different Forms of Sampling

In several natural science fields, including medicine, research is often conducted on non-human animals, including rats and monkeys. For example, chemicals that may be possible carcinogens or experimental drugs are often tested on animals for ethical or even practical reasons. The research is valuable if the underlying process is believed to be similar across species. Demonstrating that a chemical is a carcinogen in animals will often be enough to result in a ban on the chemical for human use. In other cases, research on animals is simply one piece of a program of research in which a variety of theoretical and practical concerns are examined. For example, testing drug efficacy, tolerance, and toxicity with animals is often seen as a necessary step before testing it on humans, although ethical concerns are becoming more prevalent with this practice (Goyal, 2015).

In the social and behavioral sciences, tests on non-human animals are less common. However, some research examining phenomena across different species suggests evidence of the universality of some mechanisms. For example, research may discover that memory problems in humans are associated with lower levels of a particular neurotransmitter. After this finding, researchers may prefer to test the effects of increasing levels of that neurotransmitter on animals. If these results are promising, the study may be replicated with humans. The human replication may be consistent with what was found with animals, or perhaps the results will be different, yet either result is informative.

When examining human behavior, a great deal of scientific research draws on student samples for its investigations. Within some fields in the social and behavioral sciences, such as psychology, student samples are common practice, while in others such as anthropology they are very unusual. In anthropology, and in fields which apply anthropological or ethnographic methods, the issue of generalizability tends to be less important than in many other social and behavioral sciences, often because they focus on unique populations and do not seek to generalize the findings to other populations (Gold, 1997; Honigmann, 2003). There are also some fields such as business that make frequent use of both student and non-student samples (Espinosa & Ortinau, 2016; Peterson, 2001; Simonson et al., 2001).

Importantly, for some fields, especially those in the qualitative and ethnographic tradition, research is often less about understanding the breadth of the overall population, but instead about developing a deeper understanding about a narrower

population. This is because in these cases the population of interest is not people in general, but instead only a particular group such as Kaska Indians or Samoan girls (Honigmann, 2003). This approach has even developed its own term – "sociological sampling" (Gold, 1997). Instead of breadth, the aim is increasing the validity of those data by avoiding observer bias and documenting the findings (Gold, 1997, p. 399). As a result, one basic question to ask is who you want your research to generalize to. That is, before deciding on a sample, it is important to think about what population we want to generalize to, and then think about ways to generate a sample of that population. While we focus on "generalization" to population to which the results can apply, it can also apply to the external validity of the situation and the generalized knowledge that results (Shapiro, 2002).

If we examine current practice, student samples are a common source of data in the social and behavioral sciences. Instead of dismissing this research out of hand, we should consider two questions: first, is the sample appropriate to the question being posed and, second, does the sample used raise a concern about the validity and generalizability of those findings? When making simple population estimates, a non-representative sample is indeed a problem. In theoretical investigations, would the sample that is chosen alter the underlying relationships? When looking at a hypothesized relationship between two variables, or when no inferences about the general population need to be drawn, then a student sample can be appropriate. If we are interested in understanding older adults or even all adults, then a student sample does not make sense. The mechanism underlying the process and the theory should suggest whether we can draw inferences about the general population.

Background on Student Samples

Just as the simple act of stirring a chemical solution before sampling will usually lead to a more accurate measurement (Stephan, 1948), this concept can also be applied to our understanding of sampling – we are likely to find greater reliability in a more homogeneous sample. The question of generalizability, or validity, however, is different. The issue of generalizability leads us to the critical question of whether we want to draw conclusions about people in general, and if we do, whether college students will allow us to do that. In an historical retrospective, a debate on the generalizability of student samples occurred in the field of psychology as early as the 1940s (McNemar, 1946). In 1986, David Sears suggested that reliance on college sophomores constricts our understanding of human behavior and the mind. Importantly, Sears was not suggesting that student samples were useless, but that they were limiting. In his thesis, Sears (1986) identified specific characteristics of student samples that raise concerns. The three characteristics that he identified are that students (a) may have a less strongly formulated sense of self (and a stronger need for peer approval), (b) may have higher than average cognitive skills, and (c) may have a higher level of compliance to authority. Importantly, Sears considered and evaluated the possible influence of each of these factors. His concern that students may be problematic did not result in his dismissal of student samples out

of hand, or a rejection of all findings to date; instead, he examined the potential biases themselves.

Consider the potential effects of these three characteristics if our sample were limited to students. First, if students have a weaker sense of self, students' opinions may be more volatile than those of older adults, so studies on topics such as attitude change may be biased. This does not mean that research using students cannot contribute to knowledge about attitude change; if students' opinions are more volatile, the process of opinion formation and change may not be qualitatively different but may change more readily. As a result, drawing inferences about *how* attitudes change may be correct, but the *likelihood* or *rate* of attitude change based on studies of students may be biased. Importantly, comparing persuasion results from a college student sample with those of adults is necessary to gain insights into that question.

Sears's second concern was that students have higher than average cognitive skills. It is possible that these increased skills change the nature of the mental processes that are performed. Research investigating decision-making, for example, might be biased. However, higher cognitive abilities could also work in our favor, resulting in more accurate questionnaire responses and reports of their mental activity. Interestingly, however, most studies of students show that these cognitively skilled students are mostly unsystematic in their decisions (Kahneman et al., 1982). If cognitively skilled students are unsystematic, then the public may be even more unsystematic. Findings from this line of research, therefore, might be biased in overestimating the rate of systematic thinking by the general population, but likely not our understanding of the underlying process. If the goal of the research is to look at the likelihood that this occurs, then the results are problematic; if the goal is to measure the effects of unsystematic thinking on the decisions that are made, then the results are most likely valid (for a discussion of validity, see Volume 1 of this Handbook).

The third concern Sears raised about students is that they may be more compliant to authority. Although there is some evidence of that students may be more influenced by demand effects, such as the willingness to complete a study, their compliance with our predictions is not always present (Nichols & Maner, 2008). So, let's examine the potential effect on research findings. At its most fundamental, increased demand effects may work in favor of science, making students more likely to complete the studies and complete them accurately. There are, however, two ways we typically avoid potential compliance effects in our research. First, the research subject is often "blind" to the hypothesis so the notion of artificially supporting the hypothesis should be reduced. Second, quantitative research often relies on objective measurements when possible, and also tries to separate these measurements from one another to reduce the "carryover" effects which might either color the results or allow participants to guess the hypothesis.

The points so far should not suggest that student samples are never a problem. A representative sample is usually better than a non-representative one. But not all studies need to make use of representative samples of the population, for a variety of reasons. Researchers and critics need to examine their research question and

methods to ask whether there is any reason to suspect that the findings from a student or non-representative sample may bias the results from the question being examined or the population one is trying to understand. It would be wrong to simply reject all student samples out of hand; instead, we should question whether there is any logic or evidence to believe that the results of a study using students would be different if they were obtained from a broader sample (Greenberg, 1987). This should be done with any sample.

Some critics seem to suggest that there is no way of knowing in which instances student samples may not be generalizable to the overall population. This position is then extended to propose that all studies should avoid using students altogether (James & Sonner, 2001). I believe this reflects a fundamental misunderstanding about the importance of the sample and the overall research process. These misunderstandings also ignore many generally accepted forms of scientific practice where student and other non-representative samples are used regularly in scientific research and ignore some practical reasons why student and non-representative samples provide value to the research. The assumption that a more representative sample is inherently superior is problematic, because other sampling biases may more directly affect the conclusions. For example, imagine trying to understand investment strategies and drawing a representative sample only from investment advisors – this would not represent everyday self-directed investors.

Several new ways of gathering data online can make convenience samples one of the most cost-efficient ways to gather data (Bello et al., 2009; this volume, Chapters 2 and 4). Online samples, as well as those of college students, may provide a very convenient and cost-efficient opportunity to understand human behavior. Thoughtful use of both types of non-random samples can provide better value for our research budgets and sponsors. This may also result in novel and unique research ideas and insights, especially from exploratory research. By making a distinction between descriptive research and pilot tests, versus those that test theories using surveys and experiments, we can know in advance whether the use of a student sample is potentially problematic. Toward this end I will categorize research in four basic types – descriptive, pilot, correlational, and experimental. Later I will examine the importance of the statistical analysis and the underlying theory.

Research Design

Descriptive Studies

Descriptive studies are "concerned with and designed only to describe the existing distribution of variables, without regard to causal or other hypotheses" (Aggarwal & Ranganathan, 2019, p. 34). Frequently reported by the news media, these studies draw a sample to understand the overall population. Importantly, a biased sample will not reflect the demographics, beliefs, or behaviors of the overall population. That is, these findings are inherently subject to sampling bias. A classic example of sampling bias is

the 1936 *Literary Digest* presidential election poll in the United States which showed that Langdon would defeat Roosevelt, which was different from the actual election result (Babbie, 1992, pp. 192–194; Traugott, 2011). Evidence suggests that the findings were biased by sampling people who owned either a telephone or an automobile, and in 1936, this included too many affluent and, therefore, Republican respondents. Although the poll might be interpreted to demonstrate that most of the telephone and car owners preferred Langdon to Roosevelt, it did not accurately reflect what percentage of voters preferred Langdon to Roosevelt. Extrapolating to the overall population was wrong, as the results of the election clearly indicated. In the example of the Langdon-Roosevelt poll, the results were reported as a simple univariate analysis of the percentage of the population favoring Langdon or Roosevelt. Because there were no statistical controls or measures of affluence, a variable frequently related to political preferences, nor of its relationship with whom they favored, we should be wary of the findings. No comparison of the sample to the population, or investigation of other potentially biasing factors was done. As this example illustrates, descriptive research which relies on a non-random sample is potentially problematic if one wants to draw inferences about the general population (Smith, 1983). When we apply the 1936 election estimates to the issue of student samples, this suggests that developing population estimates based on a student sample is likely problematic. This is the reason that many critics have raised concerns about the use of student samples for research purposes (Potter et al., 1993). However, the same concerns would be true of studying only Fox News or CNBC viewers.

Several studies have shown that students can vary from the public on a variety of dimensions (Barr & Hitt, 1986; Espinosa & Ortinau, 2016; Hanel & Vione, 2016; Lamb & Stem, 1980). For example, Espinosa and Ortinau (2016) demonstrated that students' ratings of restaurants differed from those of the public. As a result, these researchers and others have asserted that students should not be used in research to represent the overall population (James & Sonner, 2001). If the study is simply descriptive, this is a reasonable conclusion. However, although many studies have demonstrated an inconsistency between student and other samples, it is important to observe that many others have not (e.g., Clara et al., 2003). Further, Greenberg (1987) explains that finding some between-subjects differences does not demonstrate that the relationships would be different for other participants, nor negate the value of college student samples.

It is also important to observe that in descriptive research, even attempts to generate a random sample from an entire population can result in a non-representative sample (e.g., Jennings & Wlezien, 2018). Lower response rates will increase the likelihood of this occurring (Babbie, 1992, pp. 266–267). Non-representativeness can also occur via the sample itself, how it was obtained, how participants were recruited, as well as attrition through the course of research (Caspaldi & Patterson, 1987; Crabbe & Pinkerton, 1992; Dura & Kiecolt-Glaser, 1990; Edlund & Swann, 1989; Frame & Strauss, 1987; Lynch et al., 1993; Mishra et al., 1993; Norden et al., 1995; Walsch et al., 1992; Wesiner et al., 1995). The fundamental conclusion on sampling is that when we are relying on a sample it is hard to know whether we have achieved a truly "representative" view of the entire population, and this is true of student and non-student samples alike. One typical test

is to compare the demographics of the sample with the underlying population; to the extent that the sample demographics reflect the overall population, this supports the argument of having drawn a representative sample. A high response rate of above 50% is helpful but does not guarantee the results are any more representative than a lower response rate (Lesser & Kalsbeek 1992; Rindfuss et al., 2015). However, it is important to consider that even as the percentage of the public that has attended college increases and grows more diverse, a representative sample from a single country may not be representative of humans (James & Sonner, 2001).

Back to the case of the *Literary Digest* Langdon-Roosevelt poll: if the nature of the sample's possible effect on political leanings had been considered, we should have concluded that drawing a sample of people who were more affluent could bias the results. We might have even rejected the conclusion. Using a more current approach, the sample could have been "reconstituted" to correct for the oversampling of the affluent (Bowen, 1994). That is to say that the result was not necessarily "wrong" for the sample from which it was drawn; it was the interpretation of the findings to the voting population that was problematic.

Although we hope that student samples would at least provide an understanding of the student-age population, evidence suggests that even specific samples of students may not provide a good picture of other types of students, or of students in general as older students may be a more reasonable surrogate for working adults than for students (James & Sonner, 2001). So, not only is it problematic to extrapolate from students to the general public, but it is also even problematic to generalize from one sample of students to other students.

Pilot and Exploratory Research

Some research using student samples defends the use of samples as simply exploratory. The use of convenience samples for exploratory research has a long history. Stephan (1948) points out that the field of astronomy began with a focus on the most visible astronomical bodies, including the moon and the larger planets, before examining less prominent bodies. In the same way, student samples can serve a useful function for pilot studies and other exploratory research. Akin to how studies with animals can provide preliminary tests of carcinogens, research with students can be a basis for pilot and exploratory research. Even those cynical about student samples generally acknowledge the value of student samples for exploratory research (Bello et al., 2009; Potter et al., 1993). This argument acknowledges that although students may have a variety of differences from the general population, they also share many similarities.

Evidence suggests that pilot or exploratory studies that rely on student samples are useful. In the field of medicine, Casadevall and Fang (2008, p. 3836) defend the value of descriptive research, yet qualify the validity of findings with a limited sample:

> Descriptive observations play a vital role in scientific progress, particularly during the initial explorations made possible by technological breakthroughs. At its best, descriptive research can illuminate novel phenomena or give rise to novel hypotheses that can in turn be examined by hypothesis-driven research. However,

> descriptive research by itself is seldom conclusive. Thus, descriptive and hypothesis-driven research should be seen as complementary and iterative.

Bello et al. (2009, p. 363) in an editorial in the *Journal of International Business Studies* propose that "results based on students are likely to be ecologically valid if they are replicated or corroborated by results based on employees or managers." The notion that results from student samples should only be considered as preliminary until replicated with a broader or more appropriate sample are common in the literature (e.g., Ferber, 1977; James & Sonner, 2001; Potter et al., 1993; Wells, 1993). Ferber (1977), for example, suggests "One justifiable use of a convenience sample is for exploratory purposes, that is, to get different views on the dimensions of a problem, to probe for possible explanations or hypotheses, and to explore constructs for dealing with particular problems or issues." Considering those admonitions, although more than 80% of US social psychology studies use student samples, only about 5% raise generalizability as a possible limitation (Banyard & Hunt, 2000; Compeau et al., 2012). As the field of psychology has been more outspoken on this issue, the use of students is generally of less concern in other fields. For example, in the field of marketing, Ashraf and Merunka (2017) found only about 20% of studies relied on student-only samples. In the field of political science, evidence suggests that student samples are used frequently in experimental research (Krupnikov et al., 2021). Concerns about the use of student samples have been raised in criminology (Payne & Chappell, 2008) and logistics (Thomas, 2011). Although there are few studies examining the prevalence of student samples in the other behavioral sciences, there are many instances of research in these fields which compares a student sample with another type of sample, suggesting at least awareness of this concern.

Several studies have compared the results of a student sample with those of a different sample or population (e.g., Hallingberg et al., 2018). Although there are differences between students and non-students, Greenberg (1987) argues that this does not mean that any observed effects are invalid. Many studies use small-scale tests of interventions before they are attempted on a larger population and are used frequently to test new education curricula. As a preliminary test, these pilot studies may help identify issues related to a variety of interventions before they are launched on a larger scale (Beebe, 2007). Student samples have also been shown to be useful in scale construction (Pernice et al., 2008). Therefore, student samples can be useful for pilot and exploratory research because of their availability and lower costs (Henry, 2008), and may be especially useful for general exploration or as a basis for arguing for funding for a broader sample (Van Teijlingen & Hundley, 2010). Some research has provided suggestions for improving the translation from pilot to full-scale efforts (Beets et al., 2020; Hallingberg et al., 2018; Wolfe, 2013).

Correlational Studies Examining Relationships

Much research tests relationships between factors, often using surveys. The best of these studies examine predicted relationships between variables (Kardes, 1996; Lucas, 2003). In defense of this approach, many theorists propose that student

samples provide valuable insights and validity, even if tested with a constrained sample. Bello et al. (2009, p. 363) propose that, "if a study is guided by a well-defined theory with sophisticated predictions, and if the results based on student participants confirm the predictions, it is likely that these results can generalize to a target population." That is, this approach proposes that research focused on examining a relationship which depends on some underlying process, especially if it predicts a specific if-then relationship, can be tested with a convenience sample such as students. The validity of such an approach is more defensible if the underlying process or intervening variables are assessed to demonstrate its viability.

One approach to generalization argues that if the relationship holds with one population, the burden of proof may be on subsequent researchers to demonstrate the relationship does not hold for other populations; in the absence of such a demonstration, it is reasonable to accept that relationship. In support of this assertion, Heggestad et al. (2015) demonstrate that low response rates did not significantly bias their estimates of correlations between variables. Another study demonstrating consistency of underlying theoretical relationships across different samples is illustrated by Basil et al. (2002). This study examined reactions to the death of Princess Diana and compared three different samples – college students, a web-based sample, and a random-digit telephone sample. The results demonstrated significant demographic differences in the age and gender of respondents, as well as the overall level of identification with Princess Diana; however, the *relationship* between level of identification and the attitudinal and behavioral outcomes were statistically consistent. That is, the greater the identification, the greater their desire to watch her funeral, and this held across all three samples. As the authors suggest, and consistent with Bello et al.'s (2009) assertion, these findings support the notion that theoretical if-then relationships may not be as affected by sample differences as simple descriptive comparisons. Similar results exist elsewhere (e.g., Harrison, 1995).

There are other theorists who propose that constrained samples such as students often not only fail in their external validity, but also fail to demonstrate the internal validity that theory testing demands (Peterson & Merunka, 2014). To test this proposition, Peterson and Merunka (2014) analyzed studies of ethics across four dozen samples and demonstrated more variability than the Basil et al. (2002) and Heggestad et al. (2015) studies, finding significant differences not only in means, but also in variances, intercorrelations, and path parameters. Similarly, Cappelen et al. (2015), in a study comparing a game playing the role of a dictator versus one involving trust found that the student sample differed from the representative sample in the importance of moral motives, the level of selfish behavior, and the gender effects observed. These findings raise concerns about the use of student samples to test theory, which will be discussed as a third concern.

Experimental Research

A long tradition suggests that experiments are less prone to sampling issues, compared to other forms of research. This is because experiments randomly

assign participants to a condition while they directly manipulate an independent variable and measure its effect on an outcome. These two factors avoid self-selection bias. Further, participants are randomly assigned to conditions, so there is no possibility of self-selection into specific groups. For example, half are given treatment A and half treatment B. Every other factor remains the same. As a result, the only difference between participants in condition A and condition B should be the variable being manipulated (Thomas, 2011). Babbie observed that "probability sampling is seldom used in experiments to select participants from a larger population. The logic of random selection is . . . [r] andomization" (Babbie, 1992, p. 242). Because participants are randomly assigned to conditions and the researcher is manipulating the variable in question, experiments using students can provide important insights into causality and examination of underlying mechanism. In addition, because a narrower sample can reduce other sources of variation, non-representativeness may even provide a benefit (Lynch, 1982).

For these reasons, when a manipulation is randomly assigned to a group and the effect is measured, representativeness is believed to be less important (Berkowitz & Donnerstein, 1982). This approach has allowed us to demonstrate rather conclusively that the manipulation affected our sample. At the most basic, this can be referred to as within-subject differences (Greenberg, 1987). Inferential statistics then provide confidence intervals that partially reflect the likely difference in the population from which the sample was drawn. Although we cannot say definitively whether the same effect would hold for a different population, we can at least rule out self-selection bias and third variables as a possible explanation.

Lucas (2003) has a thorough discussion of the often misplaced concerns about the lack of external validity in experiments. He argues that claims that student samples reduce the external validity of experiments is wrong for four reasons. First, experiments are focused on predicted theoretical relationships. Second, few theories specify the population to which these relationships apply. Third, sampling is simply a matter of procedure, not inherent in the method. Fourth, findings always depend on the whole variety of circumstances in which they were gathered, and generalization is more related to the operationalizations, measures, and, most importantly, the accuracy of the theory. Lynch (1982) and Thomas (2011) have both argued that homogeneity in experimental research is beneficial as it will increase the likelihood of falsely rejecting the null hypothesis and therefore, he argues that homogeneous samples may have some advantages over representative samples if only a theory is false.

As one example of the lower importance of samples in experimental research, Falk et al. (2013) examined whether students were more likely to participate in a study of donations or to trust others – the self-selection problem; they found no difference in participation rates. Further, they also demonstrated similar levels of trust, thus suggesting that student samples were likely a valid means to test theories of prosocial behavior. Finding that college-age students differ on unrelated factors, such as different hobbies, interests, personality characteristics, or levels of depression,

from older adults, does not mean that the effects of an appeal for donations may not work on a different population. What is most important is the underlying nature of that response.

Analysis Method: Statistical Controls

In addition to the research design, another important factor that may influence the results of a study is the analysis method. One tool that may be applied is a more robust statistical analysis (Amaya & Presser, 2016; Krackardt, 1987). That is, as we move from a univariate to a bivariate or multivariate analysis that controls for other factors, we are potentially adjusting for other important factors. Simple univariate analysis that reports the means or percentages from the overall sample has a higher potential for bias (Amaya & Presser, 2016). Specifically, univariate statistics do not allow the researcher to control for other possible factors. Analyzing a biased sample with univariate analysis will allow whatever biases that exist in the sample to bias the results. As demonstrated in the Langdon-Roosevelt poll, simple percentages using descriptive statistics are very susceptible to bias through sampling. When considerations of income are measured, statistical analyses can appropriately adjust for these differences, and the results can be interpreted more accurately. In addition, is it possible to weight the sample by any important factors found and come up with a less biased result.

Over the past several decades, an increase in computing power has resulted in the use of more sophisticated statistical tools (Efron & Tibshirani, 1991). As a result, it is more common for researchers to make use of bivariate or multivariate statistical analyses to examine the effects of multiple independent variables – not only the ones predicted, but other third variables that can bias the results. As a result of these more sophisticated analyses, research can reveal what other factors may be biasing the outcome measures (Guttman, 1973; Meyer et al., 2019). This reduces some of the potential biases involved. For example, in the Langdon-Roosevelt poll, asking people their income or party affiliation and comparing this to the national data would have likely revealed an oversampling of rich and Republican voters. Weighting the sample accordingly, as more recent approaches to polling have done, likely would have produced a less biased result, but still can sometimes fail to accurately predict the overall outcome. Similarly, including information on students that could affect your outcome measures would allow bivariate or multivariate statistical analysis to adjust for these factors – similar to when Lynch (1982) suggests "blocking variables." Blocking variables examine other possible factors. Importantly, finding an interaction with one of your demographic factors is an indication that your results may depend on the sample and may suggest potential boundary conditions (Greenberg, 1987). Of course, this requires a broad enough sample to have a range in those demographic factors – something that may not occur with a sample of college sophomores from a traditional college or university. Statistically adjusting for sample differences is not a foolproof solution, but it does allow us to control for some of the possible contributing factors that may affect research findings.

The Importance of Theory

Previous theorists have proposed that student samples are less problematic when the research involves a test of theory. For example, "if a study is guided by a well-defined theory with sophisticated predictions, and if the results based on student participants confirm the predictions, it is likely that these results can generalize to a target population" (Bello et al., 2009, p. 363). The rationale is that evaluating specific a priori predictions reduces the likelihood of Type I errors (Coleman, 2007). Applying this reasoning, when studying a particular phenomenon with a student sample, there are two things we should do. First, we should consider whether there is reason to believe the underlying phenomenon may be different for different groups of the population. Second, we should measure any potentially relevant factors that might also affect the underlying process and outcomes.

First, the researcher should consider whether there is any reason to believe that the underlying process may be different in the sample. A researcher should consider whether the results may be affected by the sample and examine any research or theories that would shed light on any bias. Imagine that we are measuring students' response to a scary movie. Four hundred students watch a movie. Half view the scary movie; the other half watch a comedy. Not only are students compared to students, but the individuals in one condition are compared to people's reactions in another condition. Imagine that all students show higher levels of arousal response to the scary movie than the comedy. Next, this difference is shown statistically to occur by chance less than 1 time in 20. Could the researcher reasonably conclude that scary movies result in more arousal than comedies?

It might be concluded that the findings are valid, at minimum, for students. We have learned something by doing the study. We cannot say, however, whether the findings are valid for individuals beyond the group from which we sampled or the extent to which they are similar. Looking at the literature, we find that younger people generally have more robust physiological responses. However, since it is generally believed that people have similar physiology, we might conclude a similar relationship would exist for other populations. Because the underlying process of fear reactions is believed to be physiologically determined, it is therefore likely to be consistent across people. So, although the results from a younger sample may overestimate the size of the effect, the overall effect would still be expected for other populations. It may however be more difficult to attain the effect, or the effect may be smaller, in the general population, since this physiological response is expected to be larger for younger people.

If there is reason to believe that the sample would fundamentally alter the results, that researcher should broaden the sample. The theory would suggest it. In the case of responses to movies, is it reasonable to conclude that this result would hold for the general population? This is a judgment call, but if the theory is that physiological responses function similarly, it is hard to believe that something that triggers a physiological response in a 20-year-old wouldn't also trigger one in a 50-year-old. Therefore, we would expect a similar underlying process in 20-year-olds as in 50-year-olds. It seems likely, then, that we can say that people are aroused by scary

movies. This is an empirical question, but, like all empirical research, we should conclude that any association that we observed is only tentatively supported when tested with a student sample.

The second caveat is that analyses should include all factors that are likely to be relevant as variables. It would be important to measure both intervening variables suggested by the theory as well as potentially relevant demographics. After the research is completed, the researcher should consider the findings to see if the underlying process appears similar and the results varied by the demographics – for example, by examining age or gender effects. If there is evidence that the sample could have affected the results, it is incumbent on the researcher to report this. The ability to avoid possible sampling problems is important. It should make a researcher careful in thinking through which sample to use. Therefore, careful consideration of the possible influence of the sample is strongly advised in all situations. If our scary movie findings replicate, but we find that arousal varies depending on the age of the participant, this may not mean that scary movies do not elicit a similar reaction in people of different ages, but it may indicate that the absolute level of physiological responses may vary. Also, given this replication we may reasonably conclude that people of different races would react similarly, short people similar to tall people, etc. People who doubt the similarity are free to conduct a replication with the requisite sample.

Applying this test to the study of other dependent variables, an important question that would be asked is whether there is any reason to believe that students' reactions are qualitatively different from the rest of the population. As mentioned, some evidence suggests that their physiological responses may be stronger. In some cases, this may make the effects more measurable. Because of the nature of a student sample, this increased response might not bias the nature of the effect but might bias its size. However, let's imagine we are measuring attitude change. Harkening back to Sears's (1986) critique of the use of student samples, in this example it seems possible that students' attitudes are less fixed, and therefore more malleable (although this is only an assertion). In this case our test of attitude change likely would demonstrate greater attitude change than with a sample of the general population. Is there reason to believe that attitudes and the attitude change process are qualitatively different in students than in older adults? It seems unlikely, but if there is reason to suspect this difference, or any indication by looking at the interactions mentioned above, we should entertain that possibility and consider a replication with a different sample. In sum, although research results may be valid, and even reliable, replication with a variety of samples is critical if one wishes to establish the generalizability of these findings to an overall population (Deffner et al., 2022).

Recommendations for the Interpretation of Student Samples

Instead of inherently rejecting all research done with student samples, it is important to consider whether a sample could bias the results, their interpretation, or

their generalizability. For example, if a study tries to predict what people think about a current event using only univariate statistics, then a student-only sample would likely pose a threat to the validity of those findings. If, however, the study examines the relationship between political orientation and how people interpret a current event, and therefore is theoretically based, these effects are more likely to apply to the public. While we could not estimate accurately the rate of the public's political orientation from a student sample, there are fewer reasons to believe that the effects of political orientation on interpretations would be different in the overall population. The question, then, should be whether the process being examined is likely to hold for the overall population. Unless the underlying process is likely to be different, it would be wrong to reject it. If there are those who question this relationship, they are free to replicate the study using a broader sample. It is the dialogue between the theoretical advancements and the practical applications that leads to the most valuable insights in social and behavioral science. To paraphrase what is often attributed to Kurt Lewin, the only thing more valuable than a good theory may be a good theory that has been supported through a variety of different tests. Therefore, it is in our best interest to continue to use and value student samples but to be careful in suggesting to which population the results may generalize, especially when the sample could potentially bias the results or increase the possibility of confirming the hypothesis.

In addition to our concerns about the participants sampled, if we take a broader view on the issue of sampling, the question of generalizability also can be seen in the use of stimuli. To what extent would the stimuli used in your study generalize to other situations? In the case of scary movies, to generalize to the population of "scary movies" it is advantageous to rely on a variety of scary movies, something that has been referred to as M>1 research (Jackson, 1992; Jackson & Jacobs, 1983). Sampling messages is a form of replication that increases our confidence in our generalizations and allows better estimates of effect sizes (Monin & Oppenheimer, 2014). Therefore, in the same way that sampling a wider variety of participants adds greater generalizability to our findings, sampling stimuli as well as conceptual replications also add greater generalizability to our findings (Crandall & Sherman, 2016).

Conclusions

Concerns about the generalizability of research results are warranted. Although many theorists believe that students introduce less error into research than a "representative" sample through higher completion rates, greater attention, and higher cognitive ability (Burnett & Dune, 1986; Lynch, 1982), the issue of how generalizable a study's results are should always be considered. Suggestions that studies drawn from a broad sample of the public are inherently better are not always true. In addition to self-selection, non-response and attrition, the context and the stimuli may all affect the representativeness of the conclusions that are drawn (Shapiro, 2002). As a result, studies from a "non-representative" sample of students

can be less biased than a broader sample of the population. The guidelines offered here – considering the research method, whether uni-, bi-, or multivariate analyses were involved, and whether we are testing a theory with a priori predictions – should provide a means to evaluate the potential validity of student samples. These questions should provide clues about where and when the breadth of the sample might threaten our findings. If these concerns arise, a replication may be in order.

This review has focused on the potential differences between a student sample and the overall population from which those students are drawn. A more macro question, however, is whether a study based on people in the United States, or North America, Europe, or a clinical population, is representative of people in general (Nielsen et al., 2017). Some of the previously mentioned studies demonstrate that what was learned from a North American sample may not generalize to people in other parts of the world (e.g., Baláž et al., 2013; Kim et al., 2018). As a result, a student sample from a particular country is theoretically no more limiting than even a representative sample from a single country in our ability to generalize to the global population (Nielsen et al., 2017).

Although some would have us eliminate students as a source of information, we should ask whether the practical advantages outweigh the additional costs. Having an easily accessible research population at a lower cost allows us to ask more questions. In addition to the specific issues of student samples that have been the focus of this review, new ways of gathering data online mean the question of limited samples has an even broader relevance. Given limited resources such as declining levels of grant money, there may be more need to rely on limited samples, including students and online samples, especially with initial research. Thoughts on how to best use these samples can provide important insights into their value. This can result in research being able to explore more novel questions and provide unique research insights from a broader research base.

The bottom line here is that the automatic dismissal of research using student samples is not warranted. Although increasing the breadth of a sample is helpful, it does not inherently increase the validity of the findings. As I have proposed earlier (Basil, 1996, p. 439), "the hallmark of science is not the quality of the sample, but the testing of a theory in situations that allow its possible falsification." In sum, student samples are no worse than any other convenience sample and any one sample is only part of the evidence for a theory. However, we should be careful in claiming to what populations our results may generalize. Evaluating our samples with this perspective should help us understand where there might be concerns, or when the conclusions should be tempered. With this insight, researchers can discover which studies are worthy of replication. Such replications could provide answers to these concerns. The additional research should provide work for meta-analyses for years and perhaps even finally answer the questions about when we need to be concerned about the validity of student samples. Returning the discussion to Sears (1986, p. 527): "We have . . . learned a great deal from studying college sophomores in the laboratory. But it may be appropriate to be somewhat more tentative about the portrait of human nature we have developed from this database."

References

Amaya, A., & Presser, S. (2016). Nonresponse bias for univariate and multivariate estimates of social activities and roles. *Public Opinion Quarterly*, *81*(1), 1–36.

Aggarwal, R., & Ranganathan, P. (2019). Study designs: Part 2 – descriptive studies. *Perspectives in Clinical Research*, *10*(1), 34–36.

Ashraf, R., & Merunka, D. (2017). The use and misuse of student samples: An empirical investigation of European marketing research. *Journal of Consumer Behaviour*, *16* (4), 295–308.

Babbie, E. (1992). *The Practice of Social Research*, 6th ed. Wadsworth.

Baláž, V., Bačová, V., Drobná, E., Dudeková, K., & Adamík, K. (2013). Testing prospect theory parameters. *Ekonomicky časopis*, *61*, 655–671.

Banyard, P., & Hunt, N. (2000). Reporting research: Something missing? *The Psychologist: Bulletin of the British Psychological Society*, *13*(2), 68–71.

Barr, S. H., & Hitt, M. A. (1986). A comparison of selection decision models in manager versus student samples. *Personnel Psychology*, *39*(3), 599–617.

Basil, M. D. (1996). The use of student samples in communication research. *Journal of Broadcasting and Electronic Media*, *40*, 431–440.

Basil, M. D., Brown, W. J., & Bocarnea, M. C. (2002). Differences in univariate values versus multivariate relationships: Findings from a study of Diana, Princess of Wales. *Human Communication Research*, *28*, 501–514.

Beebe, L. H. (2007). What can we learn from pilot studies? *Perspectives in Psychiatric Care*, *43*(4), 213–218.

Beets, M. W., Weaver, R. G., Ioannidis, J., Geraci, M., Brazendale, K., Decker, L., & Milat, A. J. (2020). Identification and evaluation of risk of generalizability biases in pilot versus efficacy/effectiveness trials: A systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, *17*(1), 1–20.

Bello, D., Leung, K., Radebaugh, L., Tung, R. L., & Van Witteloostuijn, A. (2009). From the editors: Student samples in international business research. *Journal of International Business Studies*, *40*(3), 361–364.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, *37*(3), 245.

Bowen, G. L. (1994). Estimating the reduction in nonresponse bias from using a mail survey as a backup for nonrespondents to a telephone interview survey. *Research on Social Work Practice*, *4*, 115–128.

Burnett, J. J., & Dune, P. M. (1986). An appraisal of the use of student subjects in marketing research. *Journal of Business Research*, *14*(4), 329–343.

Cappelen, A. W., Nygaard, K., Sørensen, E. Ø., & Tungodden, B. (2015). Social preferences in the lab: A comparison of students and a representative population. *Scandinavian Journal of Economics*, *117*(4), 1306–1326.

Casadevall, A., & Fang, F. C. (2008). Descriptive science. *Infection and Immunity*, *76*(9), 3835–3836.

Caspaldi, D., & Patterson, G. R. (1987). An approach to the problem of recruitment and retention rates for longitudinal research. *Behavioral Assessment*, *9*, 169–177.

Clara, I. P., Cox, B. J., Enns, M. W., Murray, L. T., & Torgrudc, L. J. (2003). Confirmatory factor analysis of the multidimensional scale of perceived social support in clinically distressed and student samples. *Journal of Personality Assessment*, *81*(3), 265–270.

Coleman, S. (2007). Testing theories with qualitative and quantitative predictions. *European Political Science*, *6*(2), 124–133.

Compeau, D., Marcolin, B., Kelley, H., & Higgins, C. (2012). Research commentary – Generalizability of information systems research using student subjects – A reflection on our practices and recommendations for future research. *Information Systems Research*, *23*(4), 1093–1109.

Crabbe, B. D., & Pinkerton, K. A. (1992). Sources of bias in Health Commission and tobacco industry surveys in Australia. *Australian Psychologist*, *27*, 103–108.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99.

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science*, *5*(3), 25152459221106366.

Dura, J. R., & Kiecolt-Glaser, J. K. (1990). Sample bias in caregiving research. *Journals of Gerontology*, *45*, P200–P204.

Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin*, *66*(6), 485–487.

Edlund, M. J., & Swann, A. C. (1989). Continuing in treatment as a form of selection bias. *American Journal of Psychiatry*, *146*, 254–256.

Efron, B., & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science*, *253* (5018), 390–395.

Espinosa, J. A., & Ortinau, D. J. (2016). Debunking legendary beliefs about student samples in marketing research. *Journal of Business Research*, *69*(8), 3149–3158.

Falk, A., Meier, S., & Zehnder, C. (2013). Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association*, *11*(4), 839–852.

Ferber, R. (1977). Research by convenience. *Journal of Consumer Research*, *4*(1), 57–58.

Fienberg, S. E., & Tanur, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review/ Revue Internationale de Statistique*, *64*(3), 237–253.

Frame, C. L., & Strauss, C. C. (1987). Parental informed consent and sample bias in grade-school children. *Journal of Social and Clinical Psychology*, *5*, 227–236.

Gold, R. L. (1997). The ethnographic method in sociology. *Qualitative Inquiry*, *3*(4), 388–402.

Goyal, R. (2015). Animal testing in the history of anesthesia: Now and then, some stories, some facts. *Journal of Anaesthesiology, Clinical Pharmacology*, *31*(2), 149–151.

Greenberg, J. (1987). The college sophomore as guinea pig: Setting the record straight. *Academy of Management Review*, *12*(1), 157–159.

Guttman, I. (1973). Care and handling of univariate or multivariate outliers in detecting spuriosity – a Bayesian approach. *Technometrics*, *15*(4), 723–738.

Hallingberg, B., Turley, R., Segrott, J., et al. (2018). Exploratory studies to decide whether and how to proceed with full-scale evaluations of public health interventions: A systematic review of guidance. *Pilot and feasibility studies*, *4*(1), 1–12.

Hanel, P. H., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PLOS ONE*, *11*(12), e0168354.

Harrison, D. A. (1995). Volunteer motivation and attendance decisions: Competitive theory testing in multiple samples from a homeless shelter. *Journal of Applied Psychology*, *80*(3), 371–385.

Heggestad, E. D., Rogelberg, S., Goh, A., & Oswald, F. L. (2015). Considering the effects of nonresponse on correlations between surveyed variables. *Journal of Personnel Psychology*, *14*(2), 91–103.

Henry, P. J. (2008). Student sampling as a theoretical problem. *Psychological Inquiry*, *19*(2), 114–126

Honigmann, J. J. (2003). Sampling in ethnographic fieldwork. In R. G. Burgess (ed.), *Field Research: A Sourcebook and Field Manual* (pp. 134–152). Routledge.

Jackson, S. (1992). *Message Effects Research: Principles of Design and Analysis*. Guilford Press.

Jackson, S., & Jacobs, S. (1983). Generalizing about messages: Suggestions for design and analysis of experiments. *Human Communication Research*, *9*(2), 169–191.

James, W. L., & Sonner, B. S. (2001). Just say no to traditional student samples. *Journal of Advertising Research*, *41*(5), 63–71.

Jennings, W., & Wlezien, C. (2018). Election polling errors across time and space. *Nature Human Behaviour*, *2*(4), 276–283.

Kahneman, D., Slovic, R., & Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kardes, F. R. (1996). In defense of experimental consumer psychology. *Journal of Consumer Psychology*, *5*, 279–296.

Kim, H., Schimmack, U., Oishi, S., & Tsutsui, Y. (2018). Extraversion and life satisfaction: A cross-cultural examination of student and nationally representative samples. *Journal of Personality*, *86*(4), 604–618.

Kish, L. (1957). Confidence intervals for clustered samples. *American Sociological Review*, *22*(2), 154–165.

Krackardt, D. (1987). QAP partialling as a test of spuriousness. *Social Networks*, *9*(2), 171–186.

Krupnikov, Y., Nam, H. H., Style, H., Druckman, J. N., & Green, D. P. (2021). Convenience samples in political science experiments. In J. Druckman and D. Green (eds.), *Advances in Experimental Political Science* (pp. 165–183). Cambridge University Press.

Kruskal, W., & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895–1939. *International Statistical Review/Revue Internationale de Statistique*, *48*(2), pp. 169–195.

Lamb Jr., C. W., & Stem Jr., D. E. (1980). An evaluation of students as surrogates in marketing studies. *Advances in Consumer Research*, *7*(1), 796–799.

Lesser, V. M., & Kalsbeek, W. D. (1999). Nonsampling errors in environmental surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 4(4), 473–488.

Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, *21*, 236–253.

Lynch, D. L., Stern, A. E., Oates, R. K., & O'Toole, B. I. (1993). Who participates in child sexual abuse research? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *34*, 935–944.

Lynch, J. G. (1982). The role of external validity in theoretical research. *Journal of Consumer Research*, *10*, 109–111.

McNemar, Q. (1946) Opinion attitude methodology. *Psychological Bulletin*, *43*, 289–374.

Meyer, J., Kohn, I., Stahl, K., Hakala, K., Seibert, J., & Cannon, A. J. (2019). Effects of univariate and multivariate bias correction on hydrological impact projections in alpine catchments. *Hydrology and Earth System Sciences*, *23*, 1339–1354.

Mishra, S. I., Dooley, D., Catalano, R., & Serxner, S. (1993). Telephone health surveys: Potential bias from noncompletion. *American Journal of Public Health*, *83*, 94–99.

Monin, B., & Oppenheimer, D. M. (2014). The limits of direct replications and the virtues of stimulus sampling. *Social Psychology*, *45*(4), 299–300.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *Journal of General Psychology*, *135*(2), 151–166.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38.

Norden, K. A., Klein, D. N., Ferro, T., & Kasch, K. (1995). Who participates in a family study? *Comprehensive Psychiatry*, *36*, 199–206.

Payne, B. K., & Chappell, A. (2008). Using student samples in criminological research. *Journal of Criminal Justice Education*, *19*(2), 175–192.

Pernice, R. E., Ommundsen, R., Van Der Veer, K., & Larsen, K. (2008). On use of student samples for scale construction. *Psychological Reports*, *102*(2), 459–464.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 450–461.

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research*, *67*(5), 1035–1041.

Potter, W. J., Cooper, R., & Dupagne, M. (1993). The three paradigms of mass media research in mainstream communication journals. *Communication Theory*, *3*, 317–355.

Rindfuss, R. R., Choe, M. K., Tsuya, N. O., Bumpass, L. L., & Tamaki, E. (2015). Do low survey response rates bias results? Evidence from Japan. *Demographic Research*, *32*, 797–828.

Sears, D. O. (1986). College sophomore in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515–530.

Shapiro, M. A. (2002). Generalizability in communication research. *Human Communication Research*, *28*(4), 491–500.

Simonson, I., Carmon, Z., Dhar, R., Drolet, A., & Nowlis, S. M. (2001). Consumer research: In search of identity. In S. T. Fiske, D. L. Schacter, & C. Zahn-Waxler (eds.), *Annual Review of Psychology* (vol. 52, pp. 249–275). Annual Reviews.

Smith, T. M. F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society: Series A (General)*, *146*(4), 394–403.

Stephan, F. F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, *43*(241), 12–39.

Tabachnick, B. G., & Fidell, L. S. (2018). *Using Multivariate Statistics*. Pearson.

Thomas, R. W. (2011). When student samples make sense in logistics research. *Journal of Business Logistics*, *32*(3), 287–290.

Traugott, M. (2011). The accuracy of opinion polling and its relation to its future. In R. I. Shapiro & L. R. Jacobs (eds.), *The Oxford Handbook of American Public Opinion and the Media* (pp. 316–331). Oxford University Press.

Van Teijlingen, E., & Hundley, V. (2010). The importance of pilot studies. *Social Research Update*, *35*(4), 49–59.

Walsch, J. P., Sproull, L. S., & Hesse, B. W. (1992). Self-selected and randomly selected respondents in a computer network. *Public Opinion Quarterly*, *56*, 241–244.

Wells, W. D. (1993). Discovery-oriented consumer research. *Journal of Consumer Research*, *19*(4), 489–504.

Wesiner, C., Schmidt, L., & Tam, T. (1995). Assessing bias in community-based prevalence estimates: Towards an unduplicated count of problem drinkers and drug users. *Addiction*, *90*, 391–405.

Wolfe, B. E. (2013). The value of pilot studies in clinical research: A clinical translation of the research article titled "In search of an adult attachment stress provocation to measure effect on the oxytocin system." *Journal of the American Psychiatric Nurses Association*, *19*(4), 192–194.