# 1

# Probability Survey-Based Experimentation and the Balancing of Internal and External Validity Concerns[1]

*Paul J. Lavrakas[1], Courtney Kennedy[2], Edith D. de Leeuw[3], Brady T. West[4], Allyson L. Holbrook[5], and Michael W. Traugott[6]*

[1] *NORC, University of Chicago, 55 East Monroe Street, Chicago, IL 60603, USA*
[2] *Pew Research Center, Washington, DC, USA*
[3] *Department of Methodology & Statistics, Utrecht University, Utrecht, the Netherlands*
[4] *Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA*
[5] *Departments of Public Administration and Psychology and the Survey Research Laboratory, University of Illinois at Chicago, Chicago, IL, USA*
[6] *Center for Political Studies, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA*



The use of experimental designs is an extremely powerful scientific methodology for directly testing casual relationships among variables. Survey researchers reading this book will find that they have much to gain from taking greater advantage of controlled experimentation with random assignment of sampled cases to different experimental conditions. Experiments that are embedded within probability-based survey samples make for a particularly valuable research method as they combine the ability to more confidently draw causal attributions based on a true experimental design that is used with the ability to generalize the results of an experiment with a known degree of confidence to the target population which the survey has sampled (cf. Fienberg and Tanur 1987, 1989, 1996). For example, starting in the late 1980s, the rapid acceptance of using technology to gather data via computer-assisted telephone interviewing (CATI), computer-assisted personal interviewing (CAPI), and computer-assisted web interviewing (CAWI) has made it operationally easy for researchers to embed experimental designs within their surveys.

Prior to the 1990s, the history of experimentation in the social sciences, especially within psychology, essentially reflected a primary concern for maximizing the ability to identify empirically based cause-and-effect relationships – ones with strong internal validity (Campbell and Stanley 1966) – with little regard for whether the study could be generalized with confidence – i.e. to what extent the study had strong external validity (Campbell and Stanley 1966) – beyond the particular group of subjects/respondents that participated in the experiment. This latter concern remains highly pertinent with the current "replication crisis" facing the health, social, and behavioral sciences and its focus on "reproducibility" as a contemporary criterion of good experimentation (Wikipedia 2018a,b).

This is not to say that prior to 1990 that all experimental social scientists were unconcerned about external validity (cf. Orwin and Boruch 1982), but rather that the research practices of most suggested that they were not. In contrast, practical experience within the field of survey

1 The authors very much appreciate the review and valuable input received from Professor Joop Hox on an earlier version of the chapter.

*Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment,* First Edition. Edited by Paul J. Lavrakas, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, and Brady T. West.

Companion Website: www.wiley.com/go/Lavrakas/survey-research

research suggests that many survey researchers have focused on external validity, but failed to use experimental designs to enhance their research aims by increasing the internal validity of their studies. An example of this is in the realm of the political polling that is done by and for news media organizations. Far too often, political polls merely generate point estimates (e.g. 23% of the public approves of the job that the President is doing) without investigating what drives the approval and disapproval through the usage of question wording experiments (cf. Traugott and Lavrakas 2016).

A case in point: The results of a *nonexperimental* preelection study on the effects of political advertising were posted on the list-server of the American Association for Public Opinion Research (AAPOR) in 2000. This survey-based research was conducted via the Internet and reported finding that a certain type of advertising was more persuasive to potential voters than another type. By using the Internet as the data collection mode, this survey was able to display the ads – which were presented as digitized video segments – in real-time to respondents/subjects as part of the data collection process and, thereby, simulate the televised messages to which voters routinely are exposed in an election campaign. Respondents were shown all of the ads and then asked to provide answers to various questions concerning their reactions to each type of ad and its influence on their voting intentions. This was done in the individual respondent's own home in a room where the respondent normally would be watching television. Here, the Internet was used in a very effective and creative way to provide *mundane realism*[2] (and there by contribute to "ecological validity") to the research study by having survey respondents react to ads in a context quite similar to one in which they would be exposed to real political ads while they were enjoying a typical evening at home viewing television. Unlike the many social science research studies that are conducted under conditions far removed from "real life," this study went a long way toward eliminating the potential artificiality of the research environment as a serious threat to its overall validity.

Another laudable design feature of this study was that the Internet sample of respondents was chosen with a rigorous scientific sampling scheme so that it could reasonably be said to represent the population of potential American voters. The sample came from a large, randomly selected panel of U.S. households that had received Internet technology (*WebTV*) in their homes, allowing the researchers to survey subsets of the sample at various times. Unlike most social science research studies that have studied the effects of political advertising by showing the ads in a research laboratory setting (e.g. a centralized research facility on a university campus), the validity of this study was not threatened by the typical convenience sample (e.g. undergraduates "volunteering" to earn course credit) that researchers often rely upon to gather data. Thus, the results of this Internet research were based on a probability sample of U.S. households and, thereby, could reasonably be generalized to the potential U.S. electorate.

As impressive as these features of this research design were, the design had a serious, yet unnecessary, methodological limitation – one that caused it to miss a golden opportunity to add considerably to the overall validity of the conclusions that could have been drawn from its findings. The research design that was used displayed all the political ads to each respondent, one ad at a time. There were no features built into the design that controlled either for the possible effects of the order in which the respondent saw the ads or for having each respondent react to more than one ad within the same data collection session. As such, the cause-and-effect conclusions that could be drawn from this nonexperimental study design about which ads "caused" stronger respondent reactions rested on very weak footing. Since no design feature

2 Mundane realism (and ecological validity) refers to the extent to which the stimuli and procedures in a research study reflect everyday "real-life" experiences (e.g. Aronson and Carlsmith 1968; Crano and Brewer 1973; Wikipedia 2018b).

was used to control for the fact that respondents viewed multiple ads within the same data collection session, the validity of the conclusions drawn about the causality underlying the results remained little more than speculations on the part of the researchers, because such factors as the order of the ads and number of ads were not varied in a controlled manner by the researchers. Unfortunately, this missed opportunity is all too common in many survey-based research studies in the social sciences.

This study could have lent itself to the use of various experimental designs whereby a different political ad (i.e. the experimental stimuli) or different subsets of ads could have been *randomly assigned* to different subsamples of respondents. Or, the order of the presentation of the entire set of political ads could have been randomly assigned across respondents. In either case, an experimental design with random assignment would have provided the researchers with a far stronger basis (i.e. a study with greater internal validity) from which to draw causal inferences. Furthermore, such an experimental approach would have had little or no cost implications for the research budget, and under a design where one and only one ad was shown to any one respondent, would likely have saved data collection costs.

Using this example as our springboard, it is the goal of this book to explain how experimental designs *can and should* be deployed more often in survey research. It is most likely through the use of a true experimental design, with the random assignment of subjects/respondents to experimental conditions, that researchers gain the strong empirical basis from which they can make confident statements about causality. Furthermore, because of the widespread use of computer-assisted technologies in survey research, the use of a true experiment within a survey often adds little or no cost at all to the budget. In addition, embedding an experiment into a survey most often provides the advantage that the cell sizes of the different experimental groups can be much larger than in traditional experimental designs (e.g. in social psychology), which has the benefit that the statistical power of the analyses is much greater.

It is for these reasons and others that we believe that many survey researchers should utilize these powerful designs more often. And, toward that end, we believe that this consideration should be an explicit step in planning a survey, on par with traditional planning steps such as deciding upon one's sampling frame, sample size, recruitment methods, and data collection mode(s).

## 1.1 Validity Concerns in Survey Research

In their seminal book, *Experimental and Quasi-Experimental Designs for Research* (1966), Don Campbell and Julian Stanley identified four types of validity that can threaten the accuracy of any research study.

1. *Statistical conclusion validity* refers to the adequacy of the researcher's statistical methods to draw conclusions from the data in an accurate (unbiased) fashion. Without adequate statistical conclusion validity, the researchers cannot know whether they have data that demonstrate reliable relationships among their variables.
2. *Construct validity* refers to the adequacy of the measurement tools that are used for data collection to operationalize the constructs of interest (e.g. questionnaire items in a survey) in both a reliable and valid manner so that they actually are measuring what the researchers intended to measure. Without adequate construct validity, the researchers have no reliable basis from which to know what behaviors, knowledge, or attitudes actually have been measured. In survey methodology, lack of construct validity is related to *specification error and measurement errors* (cf. Biemer et al. 2004; Groves 1989).

3. *External validity* refers to the extent to which the findings of a particular study can be generalized beyond the sample from which the data were gathered, to other persons, places, or times. Without adequate external validity, researchers are hard pressed to know whether, for example, their findings are limited to merely those respondents/subjects who provided the data. In survey methodology, lack of external validity is related to *coverage error, sampling error, nonresponse error, and adjustment error* (cf. Groves 1989; Kalaian and Kasim 2008).
4. *Internal validity* refers to the extent to which the methodological design used by the researchers supports direct cause-and-effect reasoning. Without adequate internal validity, researchers may offer logical, reasoned arguments to speculate about the possible casual nature of any relationships they observe in their data, but they cannot use the internal strength of their research design itself to bolster such reasoning.

Groves' (1989) seminal work *Survey Errors and Survey Costs* provided a comprehensive and critical review of how different scientific disciplines have conceptualized validity concerns, and he correctly noted that the Campbell and Stanley framework is not embraced throughout all of social science. Groves adopted and built upon the growing tradition within survey research to utilize the *Total Survey Error* (TSE) framework to help survey researchers plan and interpret their research design. The TSE approach identifies major types of errors (bias and/or variance) inherent in survey designs including coverage error, sampling error, nonresponse error, and measurement error; see also Biemer et al. (2017) and Groves and Lyberg (2010).

1. *Coverage error* refers to error (primarily in the form of bias) that can result if one's sampling frame does not well represent one's target population. Coverage error can also result from within unit noncoverage associated with the respondent selection technique that is used (cf. Gaziano 2005).
2. *Sampling error* refers to the error, i.e. uncertainty (in the form of variance), in every sample survey that occurs merely by chance due to the fact that data are gathered from a sample of the population of interest rather than a full census being conducted. When a probability sample design is employed, the researcher can calculate the size of the sampling variance associated with the particular sample design.
3. *Nonresponse error* refers to the possible error (primarily in the form of bias) that can result if those sampled units from whom data are not gathered (e.g. not contacted or refusing households) are different in meaningful ways from those sampled units from which data are gathered (cf. Groves and Couper 1998). Nonresponse error can occur at both the unit level and the item level.
4. *Measurement error* refers to the possible error (in the form of bias and/or variance) associated with a survey's questionnaire, respondents, interviewers, and/or mode of data collection (cf. Biemer et al. 2004).

In creating this book, we embraced both the Campbell and Stanley validity framework and the TSE framework, which we view as compatible. However, many researchers who are familiar with traditional survey research methods come from disciplines that do not routinely use true random experimentation as part of their research methodologies and may be less familiar with the Campbell and Stanley framework. Thus, we expect that many using this book have not yet developed enough familiarity with the language of experimentation that is necessary to best understand what is presented in the chapters that follow. To that end, we proceed in this opening chapter with more explication about the interrelationships among the different types of validities and errors, followed by an even more detailed review of internal validity and external validity.

## 1.2 Survey Validity and Survey Error

Cook and Campbell (1979) discuss how the four forms of validity in the Campbell and Stanley framework are interrelated. Here, we also discuss how they interrelate to the phraseology and conceptualizations common to survey research.

Unless a research study utilizes statistical procedures that are adequate for the purposes to which they are deployed, analyses will be more prone to Type 1 (false positive, in finding a relationship when there actually is none) and Type 2 (false negative, in not finding a relationship when there actually is one) errors than would be the case with the use of more appropriate statistical procedures. This is especially of concern when the statistical procedures are used to test data that have been gathered in an experimental design. Type 1 errors refer to the possibility that any statistical relationship observed among two or more variables measured in a sample is a spurious one.[3] This concept is similar to sampling error: the recognition that bivariate and multivariate relationships observed between, and among, variables within a given sample of respondents may not hold up within another independent sample drawn from the same population.[4] Thus, the prudent survey researcher recognizes that the covariation that is observed between two variables within a sample may merely be due to chance variations within that particular sample and may not generalize to the entire population. Considerations of Type 1 error refer to the likelihood that an observed relationship between variables is unreliable. Within many social science disciplines, the "95% level of confidence" or threshold – i.e. probability less than 0.05 or less than 1 chance in 20 of being wrong – must be achieved or surpassed before one can conclude that the relationship is statistically "significant." This concept should be quite familiar to survey researchers. However, Type 2 error does not have its direct equivalent in traditional survey research literature. It refers to the statistical power of a research design to help researchers avoid drawing an erroneous conclusion (i.e. false negative) based on the finding that no reliable covariation exists between variables within a sample, when in fact such a relationship does exist in the population.

Related to the possibility of Type 2 error is the tradition within experimental research of not "accepting" a null hypothesis (cf. Cook et al. 1979). Instead, a prudent researcher concludes that there exists no empirical basis for accepting the alternative hypothesis if the results do not achieve statistical significance. This is more than mere semantics, as it represents the *explicit acknowledgement on the part of the researcher that any research study based on a sample, regardless of how that sample was chosen, is fallible* and that any certainty that one has about one's results within the sample is probabilistic and thus less than 100% (i.e. less than absolute certainty).[5]

Coverage error and nonresponse error are clearly related to external validity in that the three concepts address the issue of whether research findings gathered in a sample of respondents are representative of some larger, known population. Since the use of a sample, rather than a census, is almost always a survey researcher's practical preference, it should be paramount that the sample be drawn from a sampling frame that accurately "covers" the entire population of

---

3 Of note, often it is not recognized – nor understood – that the concepts of Type 1 error, Type 2 error, and sampling error are irrelevant (i.e. they do not exist) within a true census of a population. That is, those concepts apply only to research that is based upon a sample.

4 Of course, sampling error also applies to univariate sample statistics whenever the researcher's purpose is to make a point estimate on the level of a single variable within the target population. For example, 45% of a random sample of 500 likely voters may report their intention to vote for candidate A, yet when sampling error is taken into consideration this finding falls within the 95% confidence interval of 41–49%.

5 This also holds for Type 1 error. And, in addition to the effects of sampling, statistical conclusion validity refers to making correct assumptions, e.g. not assuming independence in cluster samples.

interest. However, avoiding coverage error by utilizing a representative sampling frame is not a sufficient condition for external validity. Instead, if one wants strong external validity, one also must avoid nonresponse error so that the data gathered from the respondents do not differ to a nonignorable extent from data that would have been gathered from nonrespondents (i.e. those who were sampled but from whom no data were collected due to refusals, noncontacts, etc.). In sum, the avoidance of both nonignorable coverage error and nonresponse error is a necessary condition for external validity, and together they likely comprise the sufficient condition.

Measurement error and construct validity both concern whether the tools of measurement yield reliable and valid data. Construct validity has traditionally focused on the manner in which a construct is operationalized as a specific variable. This often has meant merely how a question is worded in a survey. Measurement error within the TSE typology is a much broader concept as it explicitly includes the full data collection environment encompassing the instrument (or tool) by which the data are captured – including item wording and ordering if a questionnaire is used – and extends to the role of the interviewer when one is present. It also includes the mode of data collection and the role of the respondent herself/himself in providing reliable and valid data.

The concept of internal validity, however, especially as manifested via the use of true experimentation, is not resident within traditional survey research nomenclature. As we have noted, it is not part of the traditional steps in conceptualizing a survey research design. And, this is something we hope to change with this volume.

## 1.3 Internal Validity

Cook and Campbell (1979) define internal validity as "the approximate validity with which [one infers] that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause" (p. 37).[6] There are three conditions for establishing that a relationship between two variables ($X$ and $Y$) is a causal one, as in "$X$ causes $Y$." The researcher must demonstrate (i) covariation (that there is a reliable statistical relationship between $X$ and $Y$), (ii) temporal order ($X$ must occur before $Y$), and (iii) an attempt to eliminate other plausible explanations than changes in $X$ for observed changes in the dependent variable ($Y$). The use of a true experimental design with random assignment to conditions is of special value for the last of these three conditions, i.e. eliminating plausible alternative explanations. Figure 1.1 shows the simple formula often used to depict covariation between two variables, $X$ and $Y$, and is read, "$Y$ is a function of $X$."

$$Y = f(X)$$

**Figure 1.1** Simple mathematical representation of the interrelationship of two variables.

The concept of internal validity essentially addresses the *nature of the equal sign (=)* in the equation, i.e. is the relationship between $X$ and $Y$ a causal one? For internal validity to exist, there must be covariation demonstrated between $X$ and $Y$, and, therefore, $X$ must predict $Y$ to some statistically reliable extent. But the equal sign (and the covariation it implies) in itself does not provide evidence that cause and effect between $X$ and $Y$ has been demonstrated.

If the relationship shown in Figure 1.1 is causal, then it presupposes that $X$ causes $Y$, and thus that $X$ precedes $Y$ in a temporal sense. This is what distinguishes this specification from one

6 Cook and Campbell (1979) note that "internal validity has nothing to do with the abstract labeling of a presumed cause or effect; rather it deals with the *relationship* between the research operations *irrespective of what they theoretically represent*." (p. 38)

that says $Y$ is the cause of $X$, or that each cause the other, or that each are caused by some other unspecified variable ($Z$) – any of which could be the interpretation of the correlation between two variables. Only by the use of a controlled experiment or series of experiments can the nature and direction of these interrelationships be parceled out through the application of the hypothesized independent variable (under the control of the researcher) followed by the measurement of the dependent variable. This overcomes the typical problem of a cross-sectional survey where the independent and dependent variables are measured simultaneously.

The essential design feature of an experiment is the use of random *assignment* of subjects/respondents to different experimental conditions. The logic here is that with random assignment of respondents/subjects to different conditions (i.e. the different levels of the independent variable, $X$), all other factors will be equivalent except for the differences that the researcher directly controls in implementing the independent variable of interest. If statistically significant differences in the "average value"[7] of the dependent variable ($Y$) then are found between the randomly assigned groups, these differences then can be attributed to the independent variable that the researcher has controlled. Then the researcher usually will have a solid basis to conclude that it was the controlled differences in the independent variable ($X$) that caused the observed differences between the groups in $Y$.

A simple survey-based example of this occurs when a group of respondents are randomly assigned to two conditions: often called a "Control" condition and a "Treatment" condition. For this example, assume that a questionnaire employs a so-called "split-half" experiment whereby one randomly assigned half of respondents are exposed to the standard wording of a question (e.g. the wording used in the 2000 Census to measure whether someone is Hispanic; "Are you Spanish, Hispanic or Latino?"). The group receiving this standard wording is the Control Group. The other random half of respondents would be asked the question with some altered version of the wording (e.g. "Are you of Spanish, Hispanic or Latino origin?"). In this example, the group of respondents seeing or hearing the word "origin" in their questionnaire is the Treatment Group (sometimes called the "Experimental" Group), as they are receiving a different wording of the question to investigate what effect adding the word "origin" will have on the proportion in the Treatment Group that answers "Yes." The researcher has controlled the administration of the question wording (i.e. $X$, the independent variable) in order to learn whether the change in wording causes a change in the proportion of people who say "Yes" (i.e. $Y$, the dependent variable, whether they are Hispanic). And, this control over administration of the independent variable ($X$) is exercised via random assignment so that, in theory, nothing else is dissimilar between the two groups except for the slight change of wording between the control question and the treatment question. Thus, random assignment is the design equivalent to holding "all other things equal." Due to the strong internal validity of the experimental design in this example – which includes the expectation that there is no differential nonresponse between the control and treatment groups – the researcher can conclude with great confidence that any statistically significant difference between the two groups in the proportion that answered "Yes" to being Hispanic is associated with (i.e. caused by) the presence or absence of the word "origin" (the independent variable).[8]

---

7 In this use, "average value" generally means the mean, median, or mode for a group.

8 Of note, in fact it has been found that using the word "origin" when asking about one's Hispanic ethnicity does cause significantly more respondents to say "Yes" that they are Hispanic than when the question wording does not include "origin" (cf. Lavrakas et al. 2002, 2005).

## 1.4 Threats to Internal Validity

To better appreciate the power of an experimental design and random assignment, it is worth a brief review of various reasons why cause-and-effect inferences that are drawn from nonexperimental research designs are subject to many threats to their validity. For a much more detailed discussion of these and other reasons see Campbell and Stanley (1966) and Cook and Campbell (1979).

### 1.4.1 Selection

Too often the "selection" of the respondents (or subjects) that constitute different comparison groups turns out to be the main threat to a study's internal validity. For example, if a survey researcher sampled two different municipalities and measured the health of residents in each community, the researcher would have no statistical or methodological grounds on which to base any attributions about whether living in one community or the other "caused" any observed differences between the average health within the respective communities. In this example, no controlled effort was built into the study to make the two groups equivalent through random assignment, and in this example that would not be possible. As such, any observed differences between the health in each community could be due to countless other reasons than the place of residence, including a host of demographic and behavioral differences between the residential populations of each community.

Thus, any time two (or more) groups have been selected for comparison via a process other than random assignment, the researchers most often will have no solid empirical grounds on which to draw causal inferences about what may have caused any observed difference between the two groups.[9] Unfortunately, this does not stop many researchers from making unfounded causal attributions. This is especially the case in the field of epidemiological research when the health of large samples of volunteer patients is tracked over time. Such a panel survey may find many significant correlations between behavior and health (e.g. eating a lot of carrots is associated with better eyesight), but this is mere covariation and the study design, with its lack of random assignment to comparison groups, provides no internal validity to support causal inferences between the measured behaviors and health.

Furthermore, any study that allows respondents to "self-select" themselves into different comparison groups will suffer from selection as a threat to its internal validity. However, this point is different from the common way that self-selection into a sample is thought of. So-called convenience samples suffer from a self-selection sampling bias, as the researcher has no means of knowing whether a larger population is represented by a self-selected sample (this is an issue that affects external validity as discussed later in this chapter). However, the researcher could legitimately build a valid experiment into a survey that uses a convenience sample, simply by randomly assigning the self-selected respondents to different comparison groups. Thus, as long as the respondents do not self-select themselves into the treatment and control groups, it does not threaten the internal validity of the experiment, even if they have self-selected themselves into the larger sample. We also note that in experiments embedded in a survey, selection also includes differences in nonresponse rates (including break-off rates) between the experimental and control groups.

---

9 There is the possibility that a researcher deployed a quasi-experimental design, one without random assignment, but one that may have design features that avoid some of the potential threats to internal validity (cf. Cook and Campbell 1979).

### 1.4.2 History

This potential threat to internal validity refers to the possibility that something other than the independent variable may have taken place between the time respondents were first exposed to the independent variable and time of the measurement of the dependent variable. If so, then a differential "history" effect may have caused any observed differences among respondent groups in the dependent variable.

To illustrate this, consider a survey of attitudes toward local police being administered in two different, yet socioeconomically similar, communities to establish a baseline (i.e. a pretest measure). Imagine that the survey found that these two communities held essentially similar "pretest" attitudes. Then imagine that in one of the two communities, local police implemented a foot patrol program putting many more police officers on neighborhood streets. After this program is implemented for several months, both communities are resurveyed (i.e. a posttest measure), and the community with the new foot patrol program is now found to hold significantly more positive attitudes than the other community.

Could the researchers conclude with confidence (i.e. with strong internal validity) that the foot patrol program *caused* the improvement in attitudes? The answer is "No" for many reasons, including that there was no way for the researcher to control for whatever else may have occurred locally between the time that the pretest and posttest surveys were conducted that may have led to the attitudes in one community to change compared to the other. For example, a major crime may have been solved in one community in the intervening period, but not in the other. Was this the cause of more positive attitudes toward the police? Or was it the foot patrols?[10] This is how the differential history of two groups can confound any interpretation of cause when a true experiment is not used.

Furthermore, even if a research study starts out as a true experiment, subsequent uncontrolled history between randomly assigned groups can undermine the experiment and, thereby, its internal validity. For example, imagine a study in which the interviewers at a survey organization were randomly assigned into two groups to be trained, separately, to administer one of two different introductory spiels to randomly selected households in order to determine the differential effects on response rates of the two introductions. If something eventful happened at one of the training sessions other than the difference in the content related to the respective introductory spiels – e.g. an interviewer and the trainer got into a heated argument about the wording of the introductory spiel thereby lowering the confidence of the rest of the interviewers in that group regarding the effectiveness of that introductory wording – then this differential "history" could pose a serious threat to the internal validity of this research study, despite it being originally designed as a true experiment. If this were to happen, then the researchers would have a weakened basis on which to conclude that it was the content of the different introductions *and only that content* that caused any observable differences in response rates between the two groups of respondents.

All this notwithstanding, in many survey-based experiments, history is not a likely threat to internal validity because the dependent variable often is measured immediately after the administration of the independent variable (e.g. most wording experiments built into a questionnaire require that the respondent answers the question immediately after being exposed to the wording), but in other instances, the researcher must be very conscious of the possibility that history may have invalidated the integrity of the experimental design.

---

10 The posttest survey in each community could be used to gather information to determine if there were any differential history effects in the communities. Although this may appear to be a prudent approach to take, it is not possible to do this in such a comprehensive manner as to be able to rule out all potential history effects.

### 1.4.3 Instrumentation

Any time a measurement instrument, e.g. a survey question, is changed between a pre- and postperiod, any observed changes in the dependent variable of interest may be due solely to the change in instrumentation as opposed to real changes between the two groups due to a treatment or stimulus. For example, take a panel survey with two waves of data collection in which all respondents were asked, "Do you support or oppose the President's new plan to reduce taxes?" in Wave 1 data collection. Suppose that after Wave 1, a random half of the respondents were exposed to a direct mail campaign touting the popularity of the new tax plan. Suppose also that after Wave 1, the President began actively campaigning on behalf of the new tax plan and received consistently positive press coverage. After some passage of months, another wave of data is gathered from the same respondents, but using the following question, "Do you support or oppose the President's popular plan to reduce taxes?" Imagine that at Wave 2, a sizably larger proportion of respondents who were exposed to the direct mail campaign said that they supported the plan than had supported it at Wave 1 and that this increase was larger than the increase in support among the nontreatment group. Would this mean that the direct mail campaign exposure caused the apparent growth within that portion of the sample exposed to it?

The answer is "No, not necessarily," because although the small change in the wording of the measure at Wave 2 may appear innocuous – and given the positive press coverage might appear to be an appropriate wording change – the use of the word "popular" in the Wave 2 version of the questions could by itself have prompted (i.e. caused) more people to "conform" with apparent public opinion and say "Yes" to the question than otherwise would have happened had the exact Wave 1 wording been used. This could especially be true for the respondents exposed to the direct mail campaign. In particular, the treatment (the direct mail campaign) may have interacted with the wording change in the posttest question to cause the disproportionate shift in expressed support of the new tax plan among the group exposed to the mail campaign. Thus, it is possible that the change in support among the treatment group would have been no different in size than the change among the group that did not receive the direct mail campaign had the question wording not been altered.

### 1.4.4 Mortality

Imagine that an experimental test of a new remedial science curriculum is implemented so that a large random *sample* of inner-city high students is randomly *assigned* to a treatment group or a control group. The control group does not receive the remedial curriculum. The treatment group receives the remedial instruction during special 30-minute class sessions held only for them at the end of their regular school days. After six months of being exposed to the remedial curriculum, researchers find that the treatment group actually scores lower in science knowledge than does the control group. Does this mean that the curriculum actually caused the treatment group to do more poorly on their science knowledge test?

Although that is possible, imagine instead that receiving the remedial education curriculum caused more students in the treatment group to remain in school after six months because they were receiving the special attention. However, in the control group, more students dropped out of school during the six months, with students having the lowest knowledge of science being the most likely to drop out. In this case, differential "mortality" (i.e. differential attrition) would render the two groups no longer equivalent when the comparison was made between each group's average science knowledge score after six months. As such, researchers must guard against respondent/subject mortality threatening the internal validity of their experiment. And, even if the researchers cannot foresee and/or control against differential mortality, the possibility that

this might occur must be measured and its possible effects taken into account before one can interpret experimental results with confidence. In particular, any survey-based experiment in which the experimental treatment causes differential response rates, but the dependent variable is something other than the response rate, is subject to the effects of differential mortality.

There are other threats to the internal validity that may undermine a research design's ability to support cause-and-effect reasoning (cf. Campbell and Stanley 1966). However, by using a true experiment with random assignment, the researcher is on a much firmer ground in making valid causal attributions than without an experimental design. In addition, if threats to internal validity are suspected, using a pretest–posttest design provides additional information that can be used to test for such threats and to some extent control for them (cf. Cook et al. 1979).

## 1.5 External Validity

Now that we have explained how *random assignment* is the cornerstone of experimentation and establishment of the interval validity of a research design, it is worth clarifying a common point of confusion, namely, the difference between random assignment and *random sampling*. Random sampling is very much a cornerstone of external validity, especially when it is done within the context of a probability sampling design. In fact, the beauty and strength of high-quality survey research is that a researcher can meld *both* random assignment and random sampling, thereby having strong internal validity *and* strong external validity whenever the survey mode of data collection is appropriate for the needs and resources of the researcher.

Many researchers who use the survey mode of data collection are much more familiar with the science of sampling than they are with the science of experimentation. Although they may not have prior familiarity with the terminology of external validity, they likely are quite familiar with the principals underlying the concerns of external validity: If one wants to represent some known target population of interest accurately, then one best utilize a sampling design that (i) well represents that population via a properly chosen sampling frame, (ii) uses a random probability sampling scheme to select respondents from the frame, thereby allowing one to generalize research findings from the sample to the population with confidence and within a known degree of sampling error, and (iii) uses successful recruitment methods to achieve a final sample that is well representative (unbiased) of the initially drawn sample. As stated earlier in this chapter, "avoidance of coverage error and nonresponse error each are necessary conditions for external validity, and together they likely comprise the sufficient condition." Thus, survey researchers need to use sampling frames that fully "cover" the target population they purport to represent and need to achieve adequate cooperation from the sampled respondents that avoids nonignorable differential nonresponse and nonresponse bias.

The *linkage* between internal validity and external validity concerns whether any cause-and-effect relationship that has been observed in a research experiment can be generalized beyond the confines of the particular sample (subjects/respondents) on which the experiment was conducted. For example, the field of psychology has a long and honored history of using experimentation with strong internal validity. However, it also has the well-known (and not so honorable) history of questionable external validity for many findings related to too often using unrepresentative samples of college undergraduates (cf. Visser et al. 2013).

To achieve strong external validity in a survey, one needs to use a probability sample and avoid both coverage error and nonresponse error. As a reminder, a probability sample is one in which each element in the sampling frame has a *nonzero* and *known* probability of selection (the reader should note that having an *equal* probability of selection is not a necessary attribute of a probability sample). Coverage error is avoided if the researcher's sampling frame "covers"

(i.e. represents) the target population to which the researcher wants to generalize the study's findings. Nonresponse error is avoided as long as the group of sampled elements (i.e. respondents) chosen from the sampling frame, but from whom no data are gathered, does not differ in meaningful ways from the group of elements that are sampled and from whom data are gathered. All in all, using a probability sample that adequately covers the target population and achieves a high response rate typically has strong external validity.

## 1.6 Pairing Experimental Designs with Probability Sampling

We cannot understate the importance of the research "power" that is afforded by an experimental design in allowing a researcher to test the causal nature of the relationship between variables with confidence. We strongly encourage survey researchers to challenge themselves in thinking about how to utilize experimentation more often within their surveys so as to achieve this power, and the resulting internal validity, in advancing theory throughout the social sciences and in explaining human behaviors and cognitions more fully. As we noted earlier, this often can be done at little or no additional cost in the data collection process, and sometimes can even save costs as it may reduce the amount of data that must be gathered from any one respondent. We believe that the value of experimentation will be realized more often if survey researchers build into their survey planning process a decision stage where they explicitly consider how, if at all, to use experimentation in their survey. Similarly, we challenge social scientists who are familiar with the strengths of experimentation to improve their research studies by deploying probability sample surveys whenever those are appropriate for the topic of study. (Of note, Chapter 21 of this book provides a discussion of the usage of nonprobability sampling in survey research, including when experiments are used in the survey).

Thus, it is our sincere hope that survey researchers will heed the call of Don Campbell's (1988) vision of an "Experimenting Society" and help build it within the field of survey research. The chapters that follow provide many examples of how both strong internal validity and strong external validity can be achieved by using experimentation within surveys.

## 1.7 Some Thoughts on Conducting Experiments with Online Convenience Samples

Historically, there have been a great many social science research studies that have been conducted using nonprobability samples and that has included the vast majority of experimental studies. Reasons for this include the lower cost and less time that it requires to use a nonprobability sample for one's experiment. But it was also due to a greater concern among many experimenters about internal validity than external validity. Nowadays, many research studies are conducted using respondents/subjects from the myriad of readily available online convenience samples, which have been created using nonprobability methods. Primarily due to concerns with costs and timing, many of these experimenters seem to either assume or at least act as though, the quality of their research will not suffer by ignoring external validity concerns. Online convenience platforms include panel surveys, intercept surveys, and crowdsourced labor markets. Relative to surveys using probability-based designs, online convenience surveys tend to be substantially less expensive and faster to conduct, and thus their considerable attraction to many researchers.

Proponents of online convenience sources point to their greater representativeness relative to in-person convenience samples (e.g. Hauser and Schwarz 2016; Levay, Freese, and

Druckman 2016; Mullinix et al. 2015), which have been the modal subject pool for decades in the fields of experimental psychology and experimental social psychology, and more recently in the field of experimental political science (Berinsky, Huber, and Lenz 2012). Additionally, several studies report that estimated treatment effects in online convenience samples are similar to those observed in probability-based studies (Berinsky, Huber, and Lenz 2012; Coppock forthcoming; Mullinix et al. 2015; Weinberg, Freese, and McElhattan 2014). However, even advocates of online convenience sources note that they do not replace the need for probability-based population samples, as the latter provide a critical baseline allowing researchers to assess the conditions under which convenience samples provide useful or misleading inferences about causal attribution (Levay et al. 2016; Mullinix et al. 2015).

For researchers with limited training in survey methods, this can be difficult terrain to navigate. It may not be clear to them how convenience samples differ from probability-based ones or what steps one might take to reduce the risk of making erroneous inferences with either kind of sample. As such, and in the tradition of Donald T. Campbell, we offer what we consider to be some suggested best practices for researchers conducting experiments with any kind of survey data, but especially data coming from a nonprobability online convenience source.

1. *Consider how the nature of the sample might interact with the experiment*: Participants in a study might differ from their counterparts in the broader population in some fundamental way that is related to their participation in the survey platform. For example, online marketplace workers presumably all have access to the Internet as well as a desire to earn supplemental income from their participation in an online convenience panel. If these characteristics are related to key variables in the experiment, but are not shared with the study's target population, then such a subject pool may yield biased results. For example, if the research goal is to study the behavior of senior citizens and there is a reason to believe that the seniors in a given sample will react to the experimental stimulus differently than seniors in the population, further investigation is warranted, as is enhanced caution on behalf of the researchers in drawing conclusions about their findings. The researcher should consider embedding checks for that possibility into the survey, as well as whether other (e.g. offline) sample sources should be used to guard against this issue.
2. *Apply weighting designed to address sample skews that could bias estimates*: Many convenience survey platforms have demographic and ideological skews different from the broader population that their creators and users hope they represent. For example, online opt-in samples tend to have disproportionately high shares of adults who live alone, collect unemployment benefits, do not have children, and are low income (Pew Research Center 2016). Laborers on Mechanical Turk skew younger and more liberal than the U.S. public (Berinsky, Huber, and Lenz 2012). Weighting is an important tool for attempting to address such skews when they threaten survey estimates, as they often do. For online convenience, surveys in particular, weighting just on core demographics (e.g. gender, age, race) tends to be insufficient. Adjusting on demographics plus additional variables associated with both propensity to be interviewed and survey outcomes can help to reduce bias (e.g. Pew Research Center 2018; Schonlau, van Soest, and Kapteyn 2007; Wang et al. 2015). That said, even careful weighting does not always reduce bias on substantive variables. A meta-analysis by Tourangeau, Conrad, and Couper (2013) found that weighting removed at most up to three-fifths of the bias in online samples and that large biases on some variables persisted. Therefore, we recommend applying weights when analyzing the effects of experiments embedded in convenience surveys and contrasting results based on the weights with those results ignoring the weights entirely. Elliott and Valliant (2017) provide more technical details about the use of weights when analyzing data from nonprobability samples.

3. *Attempt to estimate variances and discuss/disclose the assumptions involved*: Professional guidance on this point has evolved in recent years, at least as it concerns using online convenience samples to make claims about public opinion (e.g. the percentage of Americans that favor privatizing social security). The early advice was for researchers to simply refrain from reporting variance estimates (e.g. a margin of sampling error) with opt-in samples. However, as Baker et al. (2013) noted that guidance ignored decades of variance estimation work by statisticians in clinical trials, evaluation research, and other domains. Arguably, it also had the inadvertent effect of encouraging survey consumers to simply ignore sampling error. Updated guidance encourages researchers working with opt-in samples to report measures of precision for their estimates (AAPOR 2015; Baker et al. 2013). In doing so, it is important to note that probability-based, simple random sample (SRS) assumptions are inappropriate. The precision of estimates from opt-in samples is a *model-based* measure, not the average deviation from the population value over all possible samples. Researchers using opt-in data should, therefore, provide/disclose a detailed description of how their underlying model was specified, its assumptions validated, and the measure(s) calculated. The AAPOR provides guidance for doing this (2016).
4. *Exercise particular caution when making inferences about non-Whites*: One acute concern is the possibility that subgroups at elevated risk of negative outcomes (e.g. on health, employment, crime victimization, etc.) also happen to be subgroups that are not well-represented in online convenience platforms. Studies documenting that online convenience samples tend to underrepresent lower income, less educated, Black, and/or Hispanic adults provide some evidence to this effect (e.g. Baker et al. 2010; Berinsky, Huber, and Lenz 2012). Critically, this problem appears to persist even after weighting. A recent study found that *weighted* online survey estimates for Blacks and Hispanics were off by over 10 percentage points on average, relative to benchmark values for these subpopulations computed from federal studies (Pew Research Center 2016). In addition, the online samples rarely yielded accurate estimates of the marginal effects of being Hispanic or Black on substantive outcomes, when controlling for other demographics. These results suggest that researchers using online nonprobability samples for their experiments are at risk of drawing erroneous conclusions about the effects associated with race and ethnicity. At the very least, when considering models for experimental effects on measures of interest in convenience samples, researchers should consider the possibility of treatment effect heterogeneity (addressed in Chapter 22 of this volume) and test interactions between indicators for non-Whites and the treatment variable of interest.
5. *Attempt to replicate findings with other sources*: If possible, experimental researchers using nonprobability platforms should attempt to replicate their experiments using random samples of the U.S. population or whatever else may be their target population similar to the program Time Sharing Experiments in the Social Sciences (Mutz 2011). If that is cost prohibitive, a potentially useful though less informative strategy would be to use a different convenience sample source. Knowing, for example, that an estimated treatment effect observed on Mechanical Turk replicated on an opt-in survey panel would provide some support for the robustness of the finding, though this assumes that the selection mechanisms and associated biases on the two platforms are different. Given that selection mechanisms for online convenience samples are neither controlled nor well documented, this strategy is far from infallible.

Although we consider these steps as a good practice, we would emphasize that they do not eliminate the risk of significant problems when using nonprobability samples, including online convenience ones, for experimentation. We note that many of these recommendations also apply to low response rate probability-based surveys, not just convenience samples. Techniques

for addressing noncoverage and/or nonresponse biases tend to be equally applicable, if not equally effective, for both probability and nonprobability experimental designs.

## 1.8 The Contents of this Book

There are 23 substantive chapters that follow in this book, and these are organized into nine topical areas, many of which are components of the TSE framework:

- Coverage
- Sampling
- Nonresponse
- Questionnaire
- Interviewers
- Special surveys
- Trend data
- Vignettes
- Analysis.

For each area, there is a brief introduction written by the coeditors responsible for the chapters in that area. Most of the chapters were chosen by the editors through a "competitive" process by which a widely publicized Call for Chapters was distributed throughout the world. A few of the chapters were invited from the authors by the editors.

In each chapter, authors were asked to state the focus of the experimental research topic they would be addressing, provide a critical literature review of past experimentation in this topic area, present at least one original case study of an experiment that they conducted within a probability-based survey, address the strengths and limitations of their case study/studies, make proscriptive recommendations for how to best carry out experimentation in this topic area, and identify areas where additional research is needed. Most of the substantive sections of the book have multiple chapters. No effort was made to reconcile redundancies across chapters or disagreements. Instead, the editors thought that it would benefit readers to learn where there is an agreement on a topic and where it does not yet exist.

Finally, we six editors would like to very much thank the 61 authors throughout the world who contributed the 23 substantive chapters in the book. We thank them most for their intellectual contributions, including showcasing in their respective chapters at least one previously unpublished survey-based experiment that they planned, conducted, and analyzed within a probability-based survey sampling design. We also would like to thank them for their considerable commitment and patience during the four years they worked closely and very cooperatively with us.

## References

American Association for Public Opinion Research (AAPOR) (2015). Code of ethics. https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx (accessed 1 March 2019).

American Association for Public Opinion Research (AAPOR) (2016). AAPOR guidance on reporting precision for nonprobability samples. https://www.aapor.org/getattachment/Education-Resources/For-Researchers/AAPOR_Guidance_Nonprob_Precision_042216.pdf.aspx (accessed 1 March 2019).

Aronson, E. and Carlsmith, J.M. (1968). Experimentation in social psychology. In: *The Handbook of Social Psychology*, 2e, vol. 2 (ed. G. Lindzey and E. Aronson), 1–79. Reading, MA: Addison-Wesley.

Baker, R., Blumberg, S.J., Brick, J.M. et al. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly* 74: 711–781.

Baker, R., Brick, M.J., Bates, N.A. et al. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1: 90–143.

Berinsky, A.J., Huber, G.A., and Lenz, G.S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis* 20: 351–368.

Biemer, P.B., Groves, R.M., Lyberg, L.E. et al. (2004). *Measurement Error in Surveys*. New York: Wiley.

Biemer, P., de Leeuw, E., Eckman, S. et al. (2017). *Total Survey Error in Practice*. Hoboken, NJ: Wiley.

Campbell, D.T. (1988). The experimenting society. In: *Methodology and Epistemology for the Social Sciences: Selected Papers of Donald T. Campbell* (ed. S. Overman). Chicago: University of Chicago Press.

Campbell, D.T. and Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cook, T.D. and Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Fields Settings*. Cambridge, MA: Houghton Mifflin.

Cook, T.D., Gruder, C.L., Hennigan, K.M., and Flay, B.R. (1979). The history of the sleeper effect: some logical pitfalls in accepting the null hypotheses. *Psychological Bulletin* 86: 662–679.

Coppock, A. (2018). Generalizing from survey experiments conducted on mechanical turk: a replication approach. *Political Science Research and Methods* 1–16.

Crano, W.D. and Brewer, M.B. (1973). *Principles of Research in Social Psychology*. New York: McGraw-Hill, Inc.

Elliott, M.R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science* 32 (2): 249–264.

Fienberg, S.E. and Tanur, J.M. (1987). Experimental and sampling structures: parallels diverging and meeting. *International Statistical Review* 55 (1): 75–96.

Fienberg, S.E. and Tanur, J.M. (1989). Experimentally combining cognitive and statistical approaches to survey design. *Science* 243: 1017–1022.

Fienberg, S.E. and Tanur, J.M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review* 64: 237–253.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly* 69: 124–157.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. and Lyberg, L. (2010). Total survey error past, present, and future. *Public Opinion Quarterly* 74 (5): 849–879.

Kalaian, S.A. and Kasim, R.M. (2008). External validity. In: *Encyclopedia of Survey Research Methods* (ed. P.J. Lavrakas), 254–256. Thousand Oaks, CA: Sage.

Hauser, D.J. and Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavioral Research* 48: 400–407.

Lavrakas, P.J., Courser, M., and Diaz-Castillo, L. (2002) Differences between Hispanic 'origin' and Hispanic 'identity' and their implications. 2002 American Association for Public Opinion Research Conference, St. Petersburg, FL.

Lavrakas, P.J., Courser, M.W., and Diaz-Castillo, L. (2005). What a difference a word can make: new research on the differences between hispanic 'origin' and hispanic 'identity' and their implications. 2005 American Association for Public Opinion Research; Miami, FL.

Levay, K.E., Freese, J., and Druckman, J.N. (2016). The demographic and political composition of mechanical turk samples. *SAGE Open* 6: 1–17.

Mullinix, K.J., Leeper, T.J., Druckman, J.N., and Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science* 2: 109–138.

Mutz, D. (2011). *Population-Based Experiments*. Princeton, NJ: Princeton University Press.

Orwin, R.G. and Boruch, R.F. (1982). RRT meets RDD: statistical strategies for assuring response privacy in telephone surveys. *Public Opinion Quarterly* 46 (4): 560–571.

Pew Research Center (2016). Evaluating Online Nonprobability Surveys. http://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys (accessed 1 March 2019).

Pew Research Center (2018). For Weighting Online Opt-in Samples, What Matters Most? http://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most (accessed 1 March 2019).

Schonlau, M., van Soest, A., and Kapteyn, A. (2007). Are "webographic" or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods* 1: 155–163.

Tourangeau, R., Conrad, F.G., and Couper, M.P. (2013). *The Science of Web Surveys*. Oxford: Oxford University Press.

Traugott, M.W. and Lavrakas, P.J. (2016). *The Voter's Guide to Election Polls*, 5e. Lulu Online Publication.

Visser, P., Krosnick, J.A., and Lavrakas, P.J. (2013). Survey research. In: *Handbook of Research Methods in Personality and Social Psychology*, 2e (ed. H.T. Reis and C.M. Judd). Cambridge: Cambridge University Press.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting* 31: 980–991.

Weinberg, J.A., Freese, J., and McElhattan, D. (2014). Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample. *Sociological Science* 1: 292–310.

Wikipedia (2018a). Replication crisis. https://en.wikipedia.org/wiki/Replication_crisis (accessed 1 March 2019).

Wikipedia (2018b). Ecological validity. https://en.wikipedia.org/wiki/Ecological_validity (accessed 1 March 2019).