7

Validity and Experimental Manipulations

In the previous chapters we have examined both experimental and nonexperimental research, taking a largely common perspective. Although we have mentioned some of the differences between experimental work and nonexperimental analysis and some of the different types of experimental research, we have generally focused on commonalities rather than distinctions. Yet, the differences in approaches can be important and many are controversial. In this part of the book we turn to these differences. Usually the controversies have to do with arguments about the validity, robustness, or generality of particular experimental designs. Thus, before we turn to the specific differences, we begin with a review of the concept of validity in research. Then we turn to particular and sometimes controversial issues in experimentation such as the location of an experiment (whether lab or field), the subjects recruited (whether subjects are students), and how the subjects are motivated (whether financial incentives are used).

7.1 Validity of Experimental Research

Suppose that we have conducted some empirical research with either experimental or nonexperimental data. We ideally want a research design that will provide us with *valid* results that are true for the population we are analyzing and *robust* results that *generalize* beyond our target population (see Definition 2.2).[1] So, for example, before we begin to study the effect of information on voting – the effect of changes in T_i on Y_i – we would like to come up with an experimental design that will give us results that meet these criteria.

[1] We use the term population here rather than data set because questions about how the data set relates to the population are some of the issues in establishing internal validity, which is discussed later.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

Although political science has become more experimental, the most controversial questions raised about experimental research have to do with the validity, robustness, or generalizability of that research for answering substantive questions in political science. Can an experimental study of the effect of information on voting give us results that are valid and robust? We spend most of this book addressing questions of validity and robustness of all empirical research. But first we need to define these terms more precisely and deal with some of the confusions in the literature. The first such confusion is over the definitions of what we mean by validity and the types of validity. The essence of the validity of empirical research is the question: "What can we believe about what we learn from the data?" Shadish et al. (2002) or SCC (recall Section 2.4.2) use the term validity as the "approximate truth" of the inference or knowledge claim. So suppose we conduct a study, either experimental or observational, of the relationship between information and voting and we ask the validity question: What do these data tell us? This definition, however, leaves unanswered the population over which one establishes the approximate truth. We have defined the data generating process (DGP) as the source for the population from which we draw the data we use in empirical research. However, usually when we engage in empirical research we consider only a subset of the population of data generated by the DGP. For example, we may want to explain voter turnout in the United States. We probably would not be interested in the data on turnout in China in such a study.

Definition 7.1 (Validity): *The approximate truth of the inference or knowledge claim.*

When we think of validity, do we mean valid with respect to the target population of the research or is it another different population of observations? Such questions have typically been divided into two separate validity issues. This simplistic view of how to refine the concept of validity is based on the early division of Campbell (1957) and is universally used by political scientists. Specifically, political scientists generally use *internal validity* to refer to how valid results are within a target population and *external validity* to refer to how valid results are for observations not part of the target population.[2] So if our data, for example, are drawn from a U.S. election,

[2] Other terms have been used by researchers to capture the issues that we address about validity in this chapter. For example, Levitt and List (2007b) use *generalizability* and others use *parallelism* as in Wilde (1981) and Smith (1982).

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

the internal validity question would ask how valid our results are from the analysis of the data for the target population of voters in that U.S. election. The external validity question would ask how valid our results are for other populations of voters in other elections, in the United States, elsewhere in the world, or in a laboratory election.

Definition 7.2 (Internal Validity): *The approximate truth of the inference or knowledge claim within a target population studied.*

Definition 7.3 (External Validity): *The approximate truth of the inference or knowledge claim for observations beyond the target population studied.*

However, this simplistic division of validity masks the complex questions involved in establishing validity and the interconnectedness between internal and external validity. Both internal and external validity are multifaceted concepts. In this chapter we explore both types of validity.

7.2 Deconstructing Internal Validity

As SCC and McGraw and Hoekstra (1994) discuss, nearly 40 years ago Campbell abandoned the simple binary division of validity into internal and external that still dominates political science. It is ironic that political scientists typically cite him as the authority when they use the simplistic terms. Cook and Campbell (1979) extended validity into a typology of four concepts. SCC use this typology by incorporating clarifications suggested by Cronbach (1982). In this typology, validity is divided into four types: construct, causal, statistical, and external.[3] The first three of these types together are what political scientists think of as internal validity. By exploring

[3] Cook and Campbell (1979) called causal validity "local molar causal validity." SCC explain how the term local molar causal validity explains itself (2002, p. 54): "The word *causal* in *local molar causal validity* emphasizes that internal validity is about causal inferences, not about other types of inferences that social scientists make. The word *local* emphasizes that causal conclusions are limited to the context of the particular treatments, outcomes, times, settings, and persons studied. The word *molar* recognizes that experiments test treatments that are a complex package consisting of many components, all of which are tested as a whole within the treatment condition." SCC label local molar causal validity "internal validity" because they believe that the longer term is too unwieldy and that this is what Campbell originally viewed as internal validity. Given that many political scientists think of internal validity as whatever is left over after external validity and thus includes statistical, causal, and construct validity, we define internal validity differently from SCC. SCC also call statistical validity "statistical conclusion validity." We use the shorthand terms causal and statistical validity because they are easy to remember and capture the essence of these types of internal validity.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

how each type represents a distinct question, we can better understand the different challenges involved in determining internal validity. How empirical research, either experimental or observational, establishes the validity of two of these types, causal and construct, was the focus of the previous four chapters. But before turning to these types of validity, we address statistical validity.

Definition 7.4 (Construct Validity): *Whether the inferences from the data are valid for the theory (or constructs) the researcher is evaluating in a theory testing experiment.*

Definition 7.5 (Causal Validity): *Whether the relationships the researcher finds within the target population analyzed are causal.*

Definition 7.6 (Statistical Validity): *Whether there is a statistically significant covariance between the variables the researcher is interested in and whether the relationship is sizable.*

7.2.1 Statistical Validity

Problems of Statistical Validity

Statistical validity is defined as whether there is a statistically significant covariance between the variables the researcher is interested in and whether the relationship is sizable. Suppose we find a relationship between information and voting (i.e., between T_i and Y_i) as defined in Chapter 3; statistical validity is whether the relationship is significant and sizable. Essentially this is what is often called the estimation problem of statistical analysis. Given the assumptions the researcher has made about the variables studied in the given data set, are the estimates efficient, accurate, significant, and sizable? Is the data set representative of the target population? Although these concerns seem minor compared to other matters we address later, as any empirical researcher knows, estimation is not an open-and-shut case. What does it mean when a researcher finds that the statistical relationship is just on the edge of the standard level of significance of 5%? Many now advocate an approach that focuses on reporting the actual significance level rather than just whether a result passes a threshold. Another question involved in statistical validity is whether the statistical assumptions about the distributions of the variables are supported. Are the errors estimated correctly? Is the size of the relationship consequential or not? How do we evaluate the significance of interaction terms?

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

Estimation issues can be important and are sometimes overlooked. As we discuss in Section 5.6, a popular method of empirical research which springs from experimental reasoning is what has been called the "difference-in-differences" approach to studying the effects of state laws or state policies. Researchers compare the difference in outcomes after and before the law or the new state policy for those affected by the law or state policy to the same difference in outcomes by those who have not been affected by the law or new state policy. The researchers often use ordinary least squares (OLS) in repeated cross sections or a panel of data on individuals before and after the passage of the law or new state policy. They then use the coefficient estimated for the dummy variable in the OLS that represents whether the law applies to the given observation as an estimate of the effects of the law or policy. However, Bertrand et al. (2004) pointed out that the OLS estimations are likely to suffer from possible severe serial correlation problems which when uncorrected lead to an underestimation of the standard error in estimating the coefficient and a tendency to reject null hypotheses that the law or policy has no effect when the null hypothesis should not be rejected.

The serial correlation occurs for three reasons: (1) the researchers tend to use fairly long time series, (2) the dependent variables are typically highly positively serially correlated, and (3) the dummy variable for the existence of the law or policy changes vary little over the time period estimated. The authors propose a solution – removing the time-series dimension by dividing the data into pre- and post-intervention periods and then adjusting the standard errors for the smaller number of observations this implies. They also point out that when the number of cases is large – for example, if all 50 states are included – then the estimation is less problematic. This is just one example of how statistical validity can matter in determining whether results are valid.

Statistical Replication

Statistical replication is a powerful method of verifying the statistical validity of a study. We follow Hunter (2001) and Hamermesh (2007) in dividing replication into two types. *Statistical replication* is when a researcher uses a different sample from the same population to evaluate the same theoretical implications as in the previous study or uses the same sample but a different statistical method evaluating the same theoretical implications (which some call verification), in both cases holding the construct validity of the analysis constant. *Scientific replication* is when a researcher uses a different sample, uses a different population to evaluate the same theoretical constructs,

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

or uses the same sample or a different sample from either the same or different population focusing on different theoretical implications from those constructs. We discuss scientific replication when we address external validity.

Definition 7.7 (Statistical Replication): *When a researcher uses a different sample from the same population to evaluate the same theoretical implications as in the previous study with equivalent construct validity or uses the same sample from the same population but comparing statistical techniques to evaluate the same theoretical implications as in the previous study, again with equivalent construct validity.*

It is easy to see that statistical replication is concerned with statistical validity rather than the external validity of results. In fact, researchers working with large data sets would probably be well served to engage in cross-validation, where the researcher splits the data into N mutually exclusive, randomly chosen subsets of approximately equal size and estimates the model on each possible group of $N - 1$ subsets and assesses the model's predictive accuracy based on each left out set. Although statistical replication may seem mundane, Hamermesh presents a number of interesting situations in economics where statistical replication has led to controversy.

There are examples in political science where results have been verified and called into challenge. For instance, Altman and McDonald (2003) showed that variations in how software programs make computations can, in sophisticated data analysis, lead to different empirical results in a statistical replication. In political science, statistical replication with new samples from the same target population can also lead to different results and some controversy. For example, Green et al. (1998) replicated analyses of MacKuen et al. (1989, 1992) on macropartisanship using a larger data set from the same population, calling into question the original conclusions of the analysis.[4] Because of the possibility that statistical replication may lead to different results, many political science journals now require that authors make their data plus any other necessary information for replicating the analysis available to those who may be interested. There are, of course, a number of issues having to do with the confidentiality of different data sets and sources; nevertheless, the general perspective within political science is that efforts should be made to make replication of statistical analysis possible.

[4] See also the response by Erikson et al. (1998).

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

In terms of experimental work, replication can at times be a bit more complicated, unless it is the simple verification variety as in Imai's (2005) statistical replication of Gerber and Green's (2000) mobilization study.[5] Statistical replication that involves drawing a new sample from the same population requires that a new experiment be conducted using subjects from the same target population with the same experimental protocols. Oftentimes experimentalists do this as part of their research, conducting several independent sessions of an experiment using different samples of subjects from the same pool.

7.2.2 Causal Validity and the Identification Problem

Even if the results are statistically valid, if we want to be able to say something about the effects of causes, then we need for our results to have causal validity. Causal validity is the determination of whether the relationships the researcher finds within the target population analyzed are causal. Thus, suppose we find a relationship between information and voting behavior in an election, either observational or experimental. Establishing causal validity for that relationship would mean establishing that changes in one of the variables – we posit information – causes changes in the other variable – voting behavior. Formally, using the preceding notation, changes in T_i cause changes in Y_i. We have spent the previous four chapters exploring how a researcher establishes causal validity using either the Rubin Causal Model or the Formal Theory Approach (FTA).

A concept closely related to causal validity is the notion of *identification* in econometrics. As Manski (1995, 2003) explains, econometricians have found it useful to separate out the concerns of identifying relationships from the concerns in estimating relationships. Estimation problems have to do with statistical issues of whether, given the data set analyzed and the assumptions made about the relationship between the data set and the population, the parameters of interest are efficiently and consistently estimated, or statistical validity.[6] Manski remarks (2003, p. 12): "Statistical inference seeks to characterize how sampling variability affects the conclusions that can be drawn from samples of limited size."

[5] See also Gerber and Green's (2005) response.

[6] Consistent parameter estimates are those that, under the assumptions made about the population, converge on the true population parameters as the sample size of the data set analyzed grows without bound. Efficient estimates are loosely those that have the lowest possible variance of unbiased estimators.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

In contrast, an identification problem exists when it is problematic to establish causal inferences even if the researcher has an unlimited sample from the population. Identification problems exist in many contexts. Of particular interest to political scientists is the identification problem that occurs because we cannot observe the same individual in multiple states of the world in the DGP. For example, suppose we are interested in the causal effect of education on voting. Our population is the citizens in a particular region. We cannot simultaneously observe each citizen, both educated and uneducated. Even if we have an unlimited sample from the population, we would not be able to find such observations. We can make assumptions about the probability of being educated and the reasonableness of comparing educated citizens' choices with those of uneducated citizens (and in rare cases observe them in the two states sequentially).

This type of identification problem is often labeled a selection problem because in observational analysis individuals select their education levels; they are not manipulated by the researcher. However, the problem is more fundamental than this label suggests. The difficulty arises because counterfactual observations are impossible to observe even if education could be randomly assigned to individuals – we still cannot observe the same individual both educated and uneducated. As we discussed in earlier chapters, there are experimental designs which come close to providing pseudo-counterfactual observations, and random assignment does help one "solve" the problem under particular assumptions. But even these solutions are merely close; they do not fully capture human choices in multiple states of the world simultaneously.

7.2.3 Construct Validity

Defining Construct Validity

When some political scientists think of internal validity, particularly with respect to experiments that evaluate formal models or take place in the laboratory, they are often referring to what SCC call construct validity. Construct validity has to do with how valid the inferences of the data are for the theory (or constructs) the researcher is evaluating. Essentially, establishing construct validity is an essential part of estimating causality in FTA; in FTA the goal is to investigate causal relations within a research design that has construct validity. Thus, if we think about causal validity as establishing whether changes in T_i cause changes in Y_i, construct validity takes a broader look and asks if our empirical analysis is a valid evaluation of our theory or model about why changes in T_i cause changes in Y_i.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

In experimental research the question is whether the design of the experiment is such that the variables investigated are closely equivalent to the variables the theory is concerned with. Are those things that the theory holds constant held constant in the experiment? Are the choices before the subjects the same as the choices assumed in the theory? Do the subjects have the same information about each other and about the environment that the theory assumes? Are the subjects in the experiment from the same target population that the theory addresses? In other words, is there a close match between what the theory is about and what is happening in the manipulated DGP?

In observational studies, researchers who work with formal theoretical models think about the equations underlying the theory and the equations underlying the empirical analysis and their relationship. Is the estimated empirical model derived from or equivalent to the underlying theoretical model? If there are disconnects between the empirical model and the theoretical model, to what extent do these disconnects lead one to discard the results of the research as not being relevant to the theory? These are issues that we have already addressed extensively in Chapter 6.

Construct Validity and the Generalizability of Results

Although we group construct validity as part of internal validity, as do most political scientists, doing so misses an important aspect of construct validity that makes it more than just about a given experiment. Construct validity is also about generalization. The generalization is to a theoretical construct that ideally the researcher does not view as limited to the particular empirical analysis, but a more general theory. Because of this, being able to establish construct validity can actually help build answers to external validity questions about the theory and any analysis of the theory. As SCC argue (2002, p. 93):

> [V]alid knowledge of constructs that are involved in a study can shed light on external validity questions, especially if a well-developed theory exists that describes how various constructs and instances are related to each other. Medicine, for example, has well-developed theories for categorizing certain therapies (say, the class of drugs we call chemotherapies for cancer) and for knowing how these therapies affect patients (how they affect blood tests and survival and what their side effects are). Consequently, when a new drug meets the criteria for being called a chemotherapy, we can predict much of its likely performance before actually testing it (e.g., we can say it is likely to cause hair loss and nausea and to increase survival in patients with low tumor burdens but not advanced cases). This knowledge makes the design of new experiments easier by narrowing the scope of pertinent patients and outcomes, and it makes extrapolations about treatment effects likely to be more accurate.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

In political science this is also true when a researcher works with a well-developed theory. Results from experiments with high construct validity can help us answer more general questions than those without construct validity. For example, Wittman (1983) and Calvert (1985) demonstrate in a two-candidate model of spatial competition, if the candidates have different policy preferences independent of whether they are elected and there is uncertainty about the ideal point of the median voter in the electorate, the candidates will choose divergent policy platforms in equilibrium. However, if the candidates are certain about the location of the median voter's ideal point, then the candidates converge in equilibrium. This comparative static prediction has been supported in experiments designed to have high construct validity (see Morton, 1993).

The theoretical prediction also has implications for the relationship between factors that affect whether candidates have policy preferences (such as candidate selection mechanisms) and knowledge of voter preferences and the divergence of candidate policy positions. We can extrapolate from the theory to consider other possible relationships for future empirical investigation, such as how a change in a candidate selection mechanism that makes candidates more ideological may impact candidate policy positions. For example, we may argue that in open primaries (where all registered voters are allowed to participate in the selection of candidates) candidates are less ideological than in closed primaries (where only the voters registered in a particular party are allowed to participate in a party's primary). The theory, supported by the experimental results in one target population, would then suggest that there is more divergence between candidates in closed primaries than in open primaries. Indeed, Gerber et al. (1998) find that this new theoretical prediction is supported with data on the policy positions of congressional incumbents, a different target population. Their research shows that congressional incumbents' policy positions are closer to the estimated positions of the median voters in their districts in states with open primaries than in states with closed primaries.

Although the preceding example demonstrates how empirical research with high construct validity that supports a theory in one target population can be useful as a basis for generalizing beyond the initial empirical research to implications in other target populations, a negative result from empirical research with high construct validity can also be useful. If the empirical research shows that the theory's predictions do not hold in one target population, and the research has high construct validity, then the results from the analysis can help develop a more general and robust theory, leading again

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

to new predictions about other populations beyond the target population in the original empirical analysis.

Construct Validity and External Validity

The previous section argues that construct validity of studies allows for generalization beyond those studies. The quote from SCC suggests that studies with construct validity can shed light on external validity questions. However, we do not believe such a conclusion should be taken too far. In our opinion, construct validity is not a substitute for external validity. To see why this is the case, consider what is implied by results from studies with construct validity. Suppose a researcher finds a situation in which the empirical research is considered to have construct validity and the theory's behavioral predictions are not supported. Does that mean that we should always change the theory once we find a single negative result? Although the empirical study may be considered to have construct validity, it is unlikely that a single negative result would be seen as decisive in determining the merits of the theory. Why? This is because all theories and models are abstractions from the DGP and, therefore, all have parts that are empirically false and can be proven empirically false when confronted with some observations of the DGP.[7] The question is not whether a theory can be proven empirically false, but when empirical inconsistencies with the theory matter enough for the theory to be modified or even discarded.

Similarly, suppose a researcher, again conducting empirical research considered to have construct validity, finds that the theory's behavioral predictions *are* supported. Does that mean that we should unconditionally accept the theory? Not necessarily. In our opinion, theory evaluation in the social sciences is a cumulative process that occurs through replication and complementary studies. However, because any theory can be disproved with enough data, the evaluation of theory is not purely an empirical question. As with Fudenberg (2006), we believe that theory should be judged on Stigler's (1965) three criteria: accuracy of predictions, generality, and tractability. In conclusion, construct validity is a property of a particular empirical study. However, negative or positive results from one such empirical study with construct validity are rarely adequate even if the results are strong and robust enough to accept or reject the theory. In our opinion, as we explain later, to establish external validity of results further empirical

[7] Note that a theory is always true in a theoretical sense if it is logically consistent; that is, the results or predictions follow directly from the assumptions.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

study is required, both nonexperimental and experimental if possible, to fully evaluate the value of social science theories.

7.2.4 Summary of Internal Validity

When political scientists refer to internal validity, they are often referring to three distinct and important aspects of validity: statistical, causal, and construct. It is better if we think of these issues separately, because each involves a different type of answer and has a separate set of concerns. It is quite possible – in fact, highly likely given advances in estimation techniques – that an analysis satisfies statistical validity but not causal validity or construct validity, although in some cases advances in the study of identification problems (causal validity) have outpaced estimation procedures, as discussed by Athey and Haile (2002) and exemplified in the problems discussed earlier with difference-in-differences studies.

7.3 Deconstructing External Validity

7.3.1 External, Statistical, and Ecological Validity

In contrast to internal validity, external validity is a more widely understood idea among political scientists – if you asked an average political scientist the definition of external validity, he or she would probably give you something similar to what we have written earlier. But knowing a definition and applying it are not the same, and political scientists often apply the term external validity incorrectly even if they are aware of the definition. For example, sometimes when a political scientist claims that an experiment does not have external validity, he or she is making the claim that the result is not internally valid in a statistical sense – that the sample is not a random sample from the appropriate target population and thus the conclusions are not statistically valid for the appropriate target population. But random sampling from a target population does not mean that a result is externally valid. If a researcher draws a random sample from the U.S. population to evaluate a hypothesis, the results of the analysis are not necessarily externally valid to individuals in China. External validity has to do with generalizing to populations beyond the target population, so whether one has a random sample from the target population tells one nothing about the external validity for other populations for which one has not taken a random sample.

Other times political scientists confuse external validity with *ecological validity*. Ecological validity, however, is not about the validity of results

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

from empirical research. It is about the similarity between the environment constructed in the research and a target environment. Some experimentalists call this mundane experimental realism or contextual congruence. The experimental environment is considered ecologically valid if the methods, materials, and settings of the research are similar to the target environment. Ecological validity is similar to what Harrison and List (2004) refer to as the fieldness of an experiment. For example, an experiment on voting may enhance ecological validity by being conducted in an actual polling place, using polling place equipment and registered voters.

Definition 7.8 (Ecological Validity): *Whether the methods, materials, and settings of the research are similar to a given target environment.*

However, this may or may not enhance external validity of the results because the target environment may not generalize. For example, the polling place equipment used in one jurisdiction may be different from that used in another jurisdiction. Thus, this may actually decrease the applicability of the results to different populations that use different types of equipment. Increasing ecological validity for *one* target population does not necessarily mean that the results generalize to *another* population and setting. External validity can only be established by generalizing beyond the target population and any target environment or setting. That said, increasing ecological validity and mundane realism of an experiment may help motivate subjects. We discuss this further in Chapter 10. We also return to Harrison and List's concerns about artificiality in experiments in Section 8.2.4.

7.3.2 Establishing External Validity

Suppose a researcher has been able to successfully identify and estimate a causal inference about a target population, using either experimental or nonexperimental data. Assume, for the moment, that the researcher is not engaging in theory testing and, thus, the construct validity of the initial analysis is not relevant. How can that researcher establish that the causal inference is externally valid? Or, more precisely, is it possible to establish the external validity of a causal inference that is not based on a theoretical construct without further empirical study? Without further empirical study, a researcher can only conjecture or hypothesize that his or her result has external validity based on similar studies or assumptions about the relationship between the population initially analyzed and the new population to be considered.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

Is it different if the result validates a theoretical prediction and has construct validity? Although having construct validity helps us build a more general theory and provides evidence of a more general theory, we still cannot use theory to establish external validity. External validity can be conjectured or hypothesized based on similar studies or assumptions about population similarities about any study, experimental or nonexperimental, but the *proof* of external validity is always *empirical.* Debates about external validity in the absence of such empirical proof are debates about the similarity of a study to previous studies or population similarities, but there can never be a resolution through debate or discussion alone. Researchers would be better served by conducting more empirical studies than by debating external validity in the absence of such studies.

What sort of empirical analysis is involved in establishing external validity? A researcher simply replicates the empirical results on new populations or using new variations on the experiment in terms of settings, materials, and so on. With respect to establishing the external validity of results from theory evaluations, the researcher may also test new implications of the theory on the new populations as well as the old population. We discuss these processes later in this chapter.

Scientific Replication

Scientific replication is all about establishing external validity. It is when a researcher uses either a different sample or a different population to evaluate the same theoretical constructions with the same theoretical implications or uses the same or a different sample from either the same or a different population to evaluate different theoretical implications from these constructs. It is obviously less easily mandated by journals than statistical replication because it involves taking the same theoretical constructs and applying them to new populations or evaluating new theoretical implications or taking causal inferences based on fact searching and determining if they can be identified and estimated in a different data set. Often a researcher has used considerable effort to find, build, or create, as in an experiment, the data set for a study of a target population. Usually a researcher has sought all the data that he or she could find that was relevant and leaves establishing external validity through scientific replication to other researchers.

Definition 7.9 (Scientific Replication): *When a researcher uses a different sample, a different population to evaluate the same theoretical constructs with the same theoretical implications, or the same or a different sample from either the same or a different population to evaluate different theoretical implications from these constructs.*

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

One possible way to establish some external validity for one's own empirical results is through the use of *nonrandom holdout samples* as advocated by Keane and Wolpin (2007) and Wolpin (2007). A nonrandom holdout sample is one that differs significantly from the sample used for the estimation along a dimension over which the causal inference or theoretical prediction is expected to hold. If the empirical results from the original estimation are supported with the nonrandom holdout sample, which involves observations that are well outside the support of the original data, then the results will have more external validity along this dimension. As Keane and Wolpin remark, this procedure is often used in time-series analyses and has been used in the psychology and marketing literature. They note that such a procedure was used by McFadden (1977). McFadden estimated a random utility model of travel demand in the San Francisco Bay area before the introduction of the subway system and then compared his estimates to the actual usage after the subway was introduced. The observations after the subway was introduced composed the nonrandom holdout sample. Keane and Wolpin point out that experiments can provide an ideal opportunity for analyses with nonrandom holdout samples. One can imagine that treatments can be used as subsets of the population just as in the aforementioned cross-valuation procedure. Suppose a researcher conducts K treatments on different dimensions. Then the researcher can estimate the effects of the treatments on each of the possible groups of $K - 1$ subsets as separate target populations and then assess the predictive accuracy on the subset omitted on the dimension omitted. In this fashion, the researcher can gain some traction on the external validity of his or her results.

Definition 7.10 (Nonrandom Holdout Sample): *A nonrandom holdout sample is a sample that differs significantly from the sample used for the estimation along a dimension over which the causal inference or theoretical prediction is expected to hold.*

Although it is rare for a researcher to engage in scientific replication of his or her own research as described earlier, fortunately a lot of political science research does involve this sort of replication of the research of others. Gerber and Green's voter mobilization study was a scientific replication of the original study of Gosnell and the work of Rosenstone, as discussed previously.

Scientific replication through experimentation can occur when subjects from a different target population are used with the same experimental protocols to evaluate the same theoretical implications, or subjects from the same or different target population are used to evaluate different

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

theoretical implications sometimes with a change in experimental protocols (maintaining the same theoretical constructs). For example, Potters and van Winden (2000) replicated an experiment they had conducted previously with undergraduate students (Potters and van Winden, 1996), using lobbyists. One advantage of laboratory experiments is that usually statistical verification with different samples from the same target population can be reasonably conducted as long as researchers make publicly available detailed experimental protocols. Such explicit publicly available protocols are also required for effective scientific replications, particularly if the experimenter seeks to replicate with a sample from a new target population using the same experimental design. It is generally the norm of experimentalists in political science to provide access to these protocols for such replication. We believe this should be required of all political science experimentalists.

Stress Tests and External Validity

Recall that in Chapter 6 we referred to a type of experiment called a stress test as part of FTA. A stress test is also a way for an experimentalist to explore issues of external validity when evaluating a formal model. For example, suppose a researcher has tested a theory of legislative bargaining in the laboratory. The model is one of complete information. However, the researcher relaxes some of the information available to the subjects to determine if the behavior of the subjects will be affected. The researcher has no theoretical prediction about what will happen. If the theory's predictions hold despite this new wrinkle, then the researcher has learned that the results of the first experiment can generalize, under some circumstances, to a less than complete information environment. The experimental results are robust to this change if the theory's predictions hold. If the theory's predictions do not hold, then the experimental results are not robust.

Another example would be to conduct the same complete-information legislative bargaining theory experiment with different subject pools by conducting what is called lab-in-the-field versions of the experiment (discussed in Section 8.2.3) to determine how robust the results are to changes in who participates in the experiment. Again, if the theory's predictions hold, we say that the results are robust to this change, and vice versa. Or the experimentalist may vary the frame of the experiment – perhaps the original experiment used a neutral frame and subjects were told they were players in a game without any political context. The experimentalist could introduce a political context to the experiment by telling the subjects they are legislators and they are bargaining for ministerial positions and see if this frame difference affects the subjects' choices.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

As noted in Chapter 6, the beauty of stress tests is that the experimentalist can incorporate new features of the experimental environment on a piecemeal basis and investigate each aspect of the change in an effort to test the limits of the external robustness or validity of the results. Stress tests, then, are important tools for experimentalists to test whether their results are externally valid or robust and where in particular the robustness or validity may break down.

Analyses of Multiple Studies

Narrative and Systematic Reviews. The tendency of researchers in political science is to look for new theoretical constructs or new theoretical implications from previously evaluated constructs that then become the focus of new empirical research. Alternatively, political scientists look for new target populations to evaluate existing theoretical constructs or established causal relations. Much less often do political scientists conduct reviews of research focusing on a particular research question. Yet, such reviews can be important in establishing the external validity of empirical results. In the psychology and medical literature, these types of syntheses have become commonplace to the extent that there is now a growing literature that reports on reviews of reviews.[8] Furthermore, many of the reviews in the psychology and medical literature are quantitative in nature, using statistical methods to synthesize the results from a variety of studies, which are called meta-analysis, a term coined by Glass (1976). Researchers in the medical field also distinguish between a purely narrative review and a systematic review that includes both a narrative review and an analysis of the studies, either qualitative or quantitative. In this perspective a meta-analysis is a quantitative systematic review.

Definition 7.11 (Narrative Review): *Reviews of existing literature focusing on a particular research question.*

Definition 7.12 (Systematic Review): *A narrative review that includes either a qualitative or quantitative synthesis of the reviewed studies' results.*

Definition 7.13 (Meta-analysis): *A quantitative systematic review using statistical methods for which the researcher uses study results as the unit of observation or to construct the unit of observation.*

[8] For reviews of the literature on meta-analysis in other disciplines, see the special issue of the *International Journal of Epidemiology* in 2002, Bangert-Downs (1986), Delgado-Rodriguez (2006), Egger and Smith (1997), and Montori et al. (2003).

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

Political scientists sometimes use the term meta-analysis to refer to a literature review that is mainly narrative and qualitative. Political scientists also sometimes call a study that combines a couple of different empirical studies to address a single question, such as combining a laboratory experiment with a larger survey, a meta-analysis. Technically, neither are considered meta-analyses. In meta-analysis, usually the unit of observation is either the results of an overall study or results from distinctive parts of the study. Sometimes in meta-analysis researchers use statistical results from an overall study or distinctive parts to "approximate" data pooling (see Bangert-Downs, 1986). Other times, researchers actually pool all the data from multiple studies in cases where such data are available; such analyses are not usually considered meta-analyses but simply pooled analyses. In meta-analyses the researcher works with the reported information from the study which, of course, is secondary information, and this information serves as the basis of his or her statistical analysis. We expect that, as more political scientists begin to conduct systematic quantitative reviews as found in other disciplines, meta-analysis will have the same meaning in political science that it has in other disciplines, so we define a meta-analysis more narrowly.[9]

Definition 7.14 (Pooled Analysis): *A quantitative study that pools data from multiple studies to examine a particular research question.*

Issues in Meta-analyses. In a meta-analysis a researcher first has to decide on the criteria for including a study. Setting the criteria raises a lot of questions for the researcher. For example, suppose that the researcher is more suspect of the statistical or causal validity of some studies than others; should the researcher include all studies, but use statistics to control for these differences, or simply exclude studies with less valid results? As Bangert-Downs (1986) discussed, in psychology there has been much debate over whether low-quality studies should be included in meta-analysis – whether meta-analysis is simply "garbage in-garbage out" in such cases. Consider, for example, a meta-analysis that includes some experimental studies where

[9] A number of researchers have conducted systematic reviews that they call meta-analyses with case study data using case study methods. See, for example, Strandberg's (2008) study of the relationship between party Web sites and online electoral competition and Sager's (2006) study of policy coordination in European cities. Both of these studies use a method that has developed in case study research called qualitative comparative analysis (QCA). Because our focus in this book is on quantitative research taking an experimental approach, we do not include QCA approaches in our analysis.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

causal validity is high with some nonexperimental studies where causal validity is not as high. Is it profitable to combine such studies for a meta-analysis? Alternatively, suppose that some of the data come from an experiment where random assignment has been utilized but another data set comes from an experiment without random assignment?

Studies also vary in the types of treatments and manipulations considered. Suppose that the treatment in a study is similar to the treatments given in other studies, but distinctive; to what extent can dissimilar studies be combined in an analysis that makes theoretical sense? One of the more seminal meta-analyses in psychology is Smith and Glass's (1977) study of the effects of psychotherapy. In this study the authors combined studies of a wide variety of psychotherapy from gestalt therapy to transactional analysis. What does such research tell us when so many different types of psychotherapy are combined? This is called the "apples-and-oranges" problem of meta-analysis. We could argue that doing so provides some overall measure of the effect of psychotherapy for policy makers who are choosing whether to support such therapies in general, but then what if one particular type of psychotherapy has been studied more often than it has actually been used, or has a bigger effect than others; does that skew the implications of the analysis?

After deciding on what types of studies to include, the researcher then faces additional statistical questions. What measures from the different studies should the research compare? Should the researcher compare significance and probabilities or sizes of effects? How does the researcher deal with publication biases? That is, suppose that studies showing no results or negative results are less likely to be published. How can the reviewer find information on such studies or, in the absence of such information, control for the possibility that they exist? Or, for example, suppose that the studies differ substantially in sample sizes, which has implications for comparisons across studies. How can a researcher control for these differences? Are there statistical techniques to estimate how robust the results of the reported studies are to unreported negative results? What happens if there is statistical dependence across different output measures?

Fortunately, the statistical methods used in meta-analysis are advanced enough in medicine and in psychology that researchers in political science who would like to conduct a meta-analysis can find a large literature on the methods that have been used to address these and many other methodological concerns. There are a number of textbooks on the subject (see, e.g., Hunter and Schmidt 1990). SCC also discussed meta-analysis at length in their Chapter 13. However, given the research interests of the other

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

disciplines, sometimes their answers are not appropriate for political science questions because many of the questions in medicine and psychology focus on particular isolated treatment effects of manipulations on individuals, whereas much of political science research examines effects at both individual and group levels and the interactions between the two. Furthermore, in the other disciplines, especially in medicine, it is likely that there are many studies that examine a common treatment and can be easily placed on a common metric for quantitative analyses, whereas doing so in political science may be more problematic.

Meta-analyses in Political Science. It is not surprising to us that meta-analyses are still rare in political science, mainly because it is difficult to think of a research question that has been the subject of the large number of studies needed for the statistical assumptions necessary for good meta-analysis. To our knowledge, meta-analyses have appeared at this writing only three times in the top three journals in political science, once in the *American Political Science Review* (Lau et al., 1999), once in the *Journal of Politics* (Lau et al., 2007, which is a replication of Lau et al., 1999), and once in the *American Journal of Political Science* (Doucouliagos and Ulubasoglu, 2008). Examples of meta-analyses are more numerous in specialized journals on public opinion and political psychology.

The meta-analyses of Lau et al. (1999, 2007) are instructive of how such synthesizing can lead to a deeper understanding of empirical relationships and provide insight into the complex choices facing researchers in meta-analyses. In these two studies the authors consider the empirical evidence on the effects of negative campaign advertising and find little support for the common perception in journalist circles that negative campaign advertising increases the probabilities that voters will choose the candidates who choose this strategy. As discussed earlier, the first step in the research approach used by Lau et al. (2007) was to decide on the criteria with which to include a study in their analysis. They chose to include studies that examined both actual and hypothetical political settings in which candidates or parties competed for support. Thus, they excluded studies of negative advertising in nonpolitical settings or in nonelectoral settings, but included studies for which the candidates and parties were hypothetical. If a researcher had reanalyzed previous data, they used the latest such study; however, they included studies by different researchers using different methods that used the same data set. Lau et al. also required that the study contain variation in the tone of the ads or campaigns. They focused on both studies that examined voter responses to the ads as intermediate effects as well as

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

their main interest on direct electoral effects and broader consequences on political variables such as turnout, voters' feelings of efficacy, trust, and political mood. The authors contend that these choices reflect their goal of answering the research question as to the effects of negative advertising in election campaigns. Yet, one may easily construct a meta-analysis that takes alternative focuses and uses different criteria. Ideally a researcher should consider how their criteria matter for the results provided. Lau et al. did consider the effects of using different studies from the same data set.

The second step in Lau et al.'s analysis was to do an extensive literature search to find all the relevant studies. Beyond simply surveying the literature, they contacted researchers working in the area, considered papers presented at conferences, and so on. This is a critical step in meta-analysis because it is important to avoid the "file drawer" problem of unpublished but important studies. The third step is to determine the measure for the quantitative analysis. Lau et al. focused on what is a standard technique in the psychology and medical literature, what is called Cohen's d or the *standardized mean difference statistic*, which is simply the difference in the means of the variable of interest in the treatment of interest versus the alternative treatment (or control group) divided by the pooled standard deviation of the two groups. Formally:

$$d_i = \frac{\overline{X_i^t} - \overline{X_i^c}}{s_i}, \tag{7.1}$$

where d_i is the standardized mean difference statistic for study i, $\overline{X_i^t}$ is the mean of the treatment group in the ith study, $\overline{X_i^c}$ is the mean of the control group in the ith study, and s_i is the pooled standard deviation of the two groups.

In experiments, the d statistic is relatively easy to calculate if a researcher has knowledge of the sample sizes and the standard deviations of the two groups being compared. However, if some of the studies contain nonexperimental data and are multivariate analyses, the researcher may not be able to easily calculate these measures. Lau et al. used an approximation for d in such cases that is derived from the t statistic, suggested by Stanley and Jarrell (1989), which is called by Rosenthal and Rubin (2003) the $d_{\text{equivalent}}$. Formally:

$$d_{\text{equivalent}} = \frac{2t}{\sqrt{df}}, \tag{7.2}$$

where t is the t statistic from the multivariate regression for the independent variable of interest and df is the degrees of freedom associated with the t

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

test. In their appendix, Lau et al. (1999) describe this measure in detail. This measure, of course, assumes that the independent variable associated with the t statistic is an accurate measure of the causal effect that the meta-analysis is studying. The important implicit assumptions implied by the use of this measure are explored more expansively in Chapter 5, when we discuss how causal inferences can be estimated from nonexperimental data.

After calculating the values of d, Lau et al. also had to deal with the fact that the different data sets combine different sample sizes. A number of methods exist in the literature to adjust for sample sizes (see, e.g., Hedges and Olkin, 1985). Lau et al. (1999, 2007) used a method recommended by Hunter and Schmidt (1990) to weight for sample size differences. These weights are described in the appendix to Lau et al. (1999). The authors also adjusted their measure for reliability of the variables as recommended by Hunter and Schmidt. For those outcomes for which studies report reliability, they used that measure; for studies that did not report reliability measures, they used the mean reliability for other findings within the same dependent-variable category. Finally, the authors adjusted the data for variability in the strength of the negative advertisement "treatments."

Shadish and Haddock (1994) noted that, in cases where all the studies considered use the same outcome measure, it might make sense to use the difference between raw means as the common metric. In other cases, the researcher may not be examining mean differences at all. For example, Oosterbeek et al. (2004) conducted a meta-analysis of choices of subjects in an experimental bargaining game. The analysis was a study of the determinants of the size of proposals made in the bargaining and the probability that proposals are rejected. There was no treatment or baseline in these experiments in the traditional sense because the question of interest is the extent that subjects deviate from theoretical point predictions rather than a comparative static prediction. Because the size of the bargaining pies varied as well as the relative values, Oosterbeek et al. controlled for such differences. We discuss this study more expansively in the next chapter because the study considers the effects of different subject pools in laboratory experiments.

An alternative to the d measure is the correlation coefficient as the effect size. For example, Doucouliagos and Ulubasoglu (2008) used partial correlations as their effect size measures weighted for sample size (see discussion of this measure by Ones et al. 1993; Rosenthal and Rubin, 1978). It makes sense where the studies reviewed examine the same correlational relationship among variables. Greene (2000, p. 234) provides details on how to calculate partial correlations from regression outputs of studies.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

Dougcouliagos and Ulubascoglu controlled for variations across the studies they examined in the empirical analysis of the data.

The *d* measure assumes that the study outcome is measured continuously. If the study outcome is binary, then *d* can yield problematic effect size estimates (see Fleiss, 1994; Haddock et al., 1998). In this case the effect size can be measured by the odds ratio. Formally:

$$o_i = \frac{AD}{BC}, \tag{7.3}$$

where o_i is the odds ratio for study i, A is the frequency with which the treatment occurs and there is no effect on the outcome, B is the frequency with which the treatment occurs and there is an effect on the outcome, C is the frequency with which the treatment is absent and there is no effect on the outcome, and D is the frequency with which the treatment is absent and there is an effect on the outcome.

Clearly, all of the decisions that researchers like Lau et al. make in conducting a meta-analysis affect the validity of the results. SCC and Hunter and Schmidt discuss these issues in detail.

7.3.3 Is External Validity Possible Without Satisfying Internal Validity?

Many political scientists quickly concede that experimental research has high internal validity compared with research with observational data and they dismiss experimental research (especially laboratory experiments) as being low on external validity compared with research with observational data. Both opinions tend to understate and ignore the multitude of issues involved in establishing the internal validity of both observational and experimental research, which we have discussed in this chapter. In particular, in our view, external validity can only be established for results that have been demonstrated to be internally valid in the senses we have mentioned – statistical conclusion, causal validity, and, if the empirical study involves theory testing, construct validity. If a result is not statistically significant, cannot be established to be causal in the population originally investigated, or is estimated from an empirical study that has little relevance to the theory being evaluated, then how can it possibly be considered externally valid or robust as a causal relationship? It makes no sense to say that some empirical research is low on internal validity but high on external validity.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press

7.4 Chapter Summary

In this chapter we have reviewed the concepts of internal and external validity. Internal validity has three components – statistical, causal, and construct. In previous chapters we have considered at length the methods in which researchers use experimental and nonexperimental data to establish causal and construct validity. The main contribution of this chapter is a discussion of how a researcher establishes the external validity of his or her results. We argue that while having a high degree of construct validity facilitates our ability to generalize, the ultimate proof of external validity is always empirical. Conjectures based on whether an experiment has external validity based on the extent that the experiment has ecological validity or any of the three types of internal validity are merely conjectures that must be evaluated with data.

Many of the aspects of research that some believe add to achieving external validity (such as drawing a random sample from a target population for an experiment or designing the experimental environment to be realistic to some target environment) may increase internal validity (statistical validity in the first case, causal and construct validity by motivating subjects in the second), but they cannot tell us whether the results from the experiment are externally valid. Furthermore, it is illogical to suggest that results are externally valid if they have not been demonstrated to be internally valid.

Does it make sense for many political scientists to dismiss experiments in general or laboratory experiments in particular as not having much external validity compared to nonexperimental research? As laboratory experimentalists, obviously we disagree with such a conclusion. But because this is such a pervasive view among political scientists, we discuss some of the most important concerns political scientists have about the validity of laboratory experiments in depth in the next three chapters: Chapter 8 addresses concerns about artificiality, Chapter 9 discusses the use of undergraduates as subject pools, and Chapter 10 focuses on the motivations of subjects in laboratory experiments.

https://doi.org/10.1017/CBO9780511762888.007 Published online by Cambridge University Press