Encyclopedia of Research Design

External Validity

When an investigator wants to generalize results from a research study to a wide group of people (or a population), he or she is concerned with external validity. A set of results or conclusions **[p. 467 ↓]** from a research study that possesses external validity can be generalized to a broader group of individuals than those originally included in the study. External validity is relevant to the topic of research methods because scientific and scholarly investigations are normally conducted with an interest in generalizing findings to a larger population of individuals so that the findings can be of benefit to many and not just a few. In the next three sections, the kinds of generalizations associated with external validity are introduced, the threats to external validity are outlined, and the methods to increase the external validity of a research investigation are discussed.

Two Kinds of Generalizations

Two kinds of generalizations are often of interest to researchers of scientific and scholarly investigations: (a) generalizing research findings to a specific or target population, setting, and time frame; and (b) generalizing findings across populations, settings, and time frames. An example is provided to illustrate the difference between the two kinds. Imagine a new herbal supplement is introduced that is aimed at reducing anxiety in 25-year-old women in the United States. Suppose that a random sample of all 25-year-old women has been drawn that provides a nationally representative sample within known limits of sampling error. Imagine now that the women are randomly assigned to two conditions—one where the women consume the herbal supplement as prescribed, and the other a control group where the women unknowingly consume a sugar pill. The two conditions or groups are equivalent in terms of their representativeness of 25-year-old women. Suppose that after data analysis, the group that consumed the herbal supplement demonstrated lower anxiety than the control group as measured by a paper-and-pencil questionnaire. The investigator can generalize this finding to the average 25-year-old woman in the United States, that is, the target population of the study. Note that this finding can be generalized to the average 25-year-old woman despite possible variations in how differently women in the experimental group reacted to the supplement. For example, a closer analysis of the data might reveal that women in the experimental group who exercised regularly

reduced their anxiety more in relation to women who did not; in fact, a closer analysis might reveal that only those women who exercised regularly in addition to taking the supplement reduced their anxiety. In other words, closer data analysis could reveal that the findings do not generalize across all subpopulations of 25-year-old women (e.g., those who do not exercise) even though they do generalize to the overall target population of 25-year-old women.

The distinction between these two kinds of generalizations is useful because generalizing to specific populations is surprisingly more difficult than generalizing across populations because the former typically requires large-scale studies where participants have been selected using formal random sampling procedures. This is rarely achieved in field research, where large-scale studies pose challenges for administering treatment interventions and for high-quality measurement, and participant attrition is liable to occur systematically. Instead, the more common practice is to generalize findings from smaller studies, each with its own sample of convenience or accidental sampling (i.e., a sample that is accrued expediently for the purpose of the research but provides no guarantee that it formally represents a specific target population), across the populations, settings, and time frames associated with the smaller studies. It needs to be noted that individuals in samples of convenience may belong to the target population to which one wishes to generalize findings; however, without formal random sampling, the representativeness of the sample is questionable. According to Thomas Cook and Donald Campbell, an argument can be made for strengthening external validity by means of a greater number of smaller studies with samples of convenience than by a single large study with an initially representative sample. Given the frequency of generalizations across populations, settings, and time frames in relation to target populations, the next section reviews the threats to external validity claims associated with this type of generalization.

Threats to External Validity

To be able to generalize research findings across populations, settings, and time frames, the investigator needs to have evidence that the research findings are not unique to a single population, **[p. 468 ↓]** but rather apply to more than one population. One source for this type of evidence comes from examining statistical interactions

between variables of interest. For example, in the course of data analysis, an investigator might find that consuming an herbal supplement (experimental treatment) statistically interacts with the activity level of the women participating in the study, such that women who exercise regularly benefit more from the anxiety-reducing effects of the supplement relative to women who do not exercise regularly. What this interaction indicates is that the positive effects of the herbal supplement cannot be generalized equally to all subpopulations of 25-year-old women. The presence of a statistical interaction means that the effect of the variable of interest (i.e., consuming the herbal supplement) changes across levels of another variable (i.e., activity levels of 25-year-old women). In order to generalize the effects of the herbal supplement across subpopulations of 25-year-old women, a statistical interaction cannot be observed between the two variables of interest. Many interactions can threaten the external validity of a study. These are outlined as follows.

Participant Selection and Treatment Interaction

To generalize research findings across populations of interest, it is necessary to recruit participants in an unbiased manner. For example, when recruiting female participants to take part in an herbal supplement study, if the investigator advertises the study predominantly in health food stores and obtains the bulk of participants from this location, then the research findings may not generalize to women who do not visit health food stores. In other words, there may be something unique to those women who visit health food stores and decide to volunteer in the study that may make them more disposed to the effects of a health supplement. To counteract this potential bias, the investigator could systematically advertise the study in other kinds of food stores to test whether the selection of participants from different locations interacts with the treatment. If the statistical interaction is absent, then the investigator can be confident that the research findings are not exclusive to those women who visit health food stores and, possibly, are more susceptible to the effects of an herbal supplement than other women. Thus, recruiting participants from a variety of locations and making participation as convenient as possible should be undertaken.

SSAGE researchmethods

Setting and Treatment Interaction

Just as the selection of participants can interact with the treatment, so can the setting in which the study takes place. This type of interaction is more applicable to research studies where participants experience an intervention that could plausibly change in effect depending on the context, such as in educational research or organizational psychological investigations. However, to continue with the health supplement example, suppose the investigator requires the participants to consume the health supplement in a laboratory and not in their homes. Imagine that the health supplement produces better results when the participant ingests it at home and produces worse results when the participant ingests it in a laboratory setting. If the investigator varies the settings in the study, it is possible to test the statistical interaction between the setting in which the supplement is ingested and the herbal supplement treatment. Again, the absence of a statistical interaction between the setting and the treatment variable would indicate that the research findings can be generalized across the two settings; the presence of an interaction would indicate that the findings cannot be generalized across the settings.

History and Treatment Interaction

In some cases, the historical time in which the treatment occurs is unique and could contribute to either the presence or absence of a treatment effect. This is a potential problem because it means that whatever effect was observed cannot be generalized to other time frames. For example, suppose that the herbal supplement is taken by women during a week in which the media covers several high-profile optimistic stories about women. It is reasonable for an investigator to inquire whether the positive results of taking an herbal supplement would have been obtained during a less eventful week. One way to test for the interaction between **[p. 469 ↓]** historical occurrences and treatment is to administer the study at different time frames and to replicate the results of the study.

$SAGE researchmethods**

Methods to Increase External Validity

If one wishes to generalize research findings to target populations, it is appropriate to outline a sampling frame and select instances so that the sample is representative of the population to which one wishes to generalize within known limits of sampling error. Procedures for how to do this can be found in textbooks on sampling theory. Often, the most representative samples will be those that have been selected randomly from the population of interest. This method of *random sampling for representativeness* requires considerable resources and is often associated with large-scale studies. After participants have been randomly selected from the population, participants can then be randomly assigned to experimental groups.

Another method for increasing external validity involves *sampling for heterogeneity.* This method requires explicitly defining target categories of persons, settings, and time frames to ensure that a broad range of instances from within each category is represented in the design of the study. For example, an educational researcher interested in testing the effects of a mathematics intervention might design the study to include boys and girls from both public and private schools located in small rural towns and large metropolitan cities. The objective would then be to test whether the intervention has the same effect in all categories (e.g., whether the mathematics intervention leads to the same effect in boys and girls, public and private schools, and rural and metropolitan areas). Testing for the effect in each of the categories requires a sufficiently large sample size in each of the categories. Deliberate sampling for heterogeneity does not require random sampling at any stage in the design, so it is usually viable to implement in cases where investigators are limited by resources and in their access to participants. However, deliberate sampling does not allow one to generalize from the sample to any formally specified population. What deliberate sampling does allow one to conclude is that an effect has or has not been obtained within a specific range of categories of persons, settings, and times. In other words, one can claim that "in at least one sample of boys and girls, the mathematics intervention had the effect of increasing test scores."

There are other methods to increase external validity, such as the *impressionistic modal instance model*, where the investigator samples purposively for specific types

of instances. Using this method, the investigator specifies the category of person, setting, or time to which he or she wants to generalize and then selects an instance of each category that is impressionistically similar to the category mode. This method of selecting instances is most often used in consulting or project evaluation work where broad generalizations are not required. The most powerful method for generalizing research findings, especially if the generalization is to a target population, is random sampling for representativeness. The next most powerful method is random sampling for heterogeneity, with the method of impressionistic modal instance being the least powerful. The power of the model decreases as the natural assortment of individuals in the sample dwindles. However, practical concerns may prevent an investigator from using the most powerful method.

Jacqueline P. Leighton

See also

Further Readings

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

Tebes, J. K. External validity and scientific psychology. American Psychologist, (2000). vol. 55(12), pp. 1508–1509.

$SAGE researchmethods**