CHAPTER 2

Searching and Screening

The Practical Screen and Methodological Quality

A Reader's Guide

Purpose of This Chapter

An important activity in the search for literature is to decide on criteria for including and excluding articles. The most efficient searches use two screens to select the studies that will be reviewed. The first screen is primarily practical. You use it to identify a broad range of potentially useful studies. This chapter explains how to use typical practical screening criteria such as a study's content, publication language, research setting and methods, and funding source, as well as the type of publication in which it appears.

The second screen is for methodological quality, and it is used to narrow the search by identifying the best available studies in terms of their adherence to methods that scientists and scholars rely on to gather sound evidence. *Methodological quality* refers to how well a study has been designed and implemented to achieve its objectives. Focusing on studies that have high quality is the only guarantee you have that the results of the review will be accurate.

The highest quality studies come closest to adhering to rigorous research standards. A useful way of thinking about research standards is in terms of the quality of study design and sampling, data collection, analysis, interpretation, and reporting. Study reports should provide sufficient information about their methods so that the reviewer has no trouble distinguishing high- from low-quality research. Among the questions the reviewer needs be able to answer are these: Is the research design internally and externally valid? Are the study's data sources reliable and valid? Are the analytic methods appropriate given the characteristics and quality of the study's data? Are the results meaningful in practical and statistical terms? Are the results presented in a cogent manner, describing the study's strengths and weaknesses?

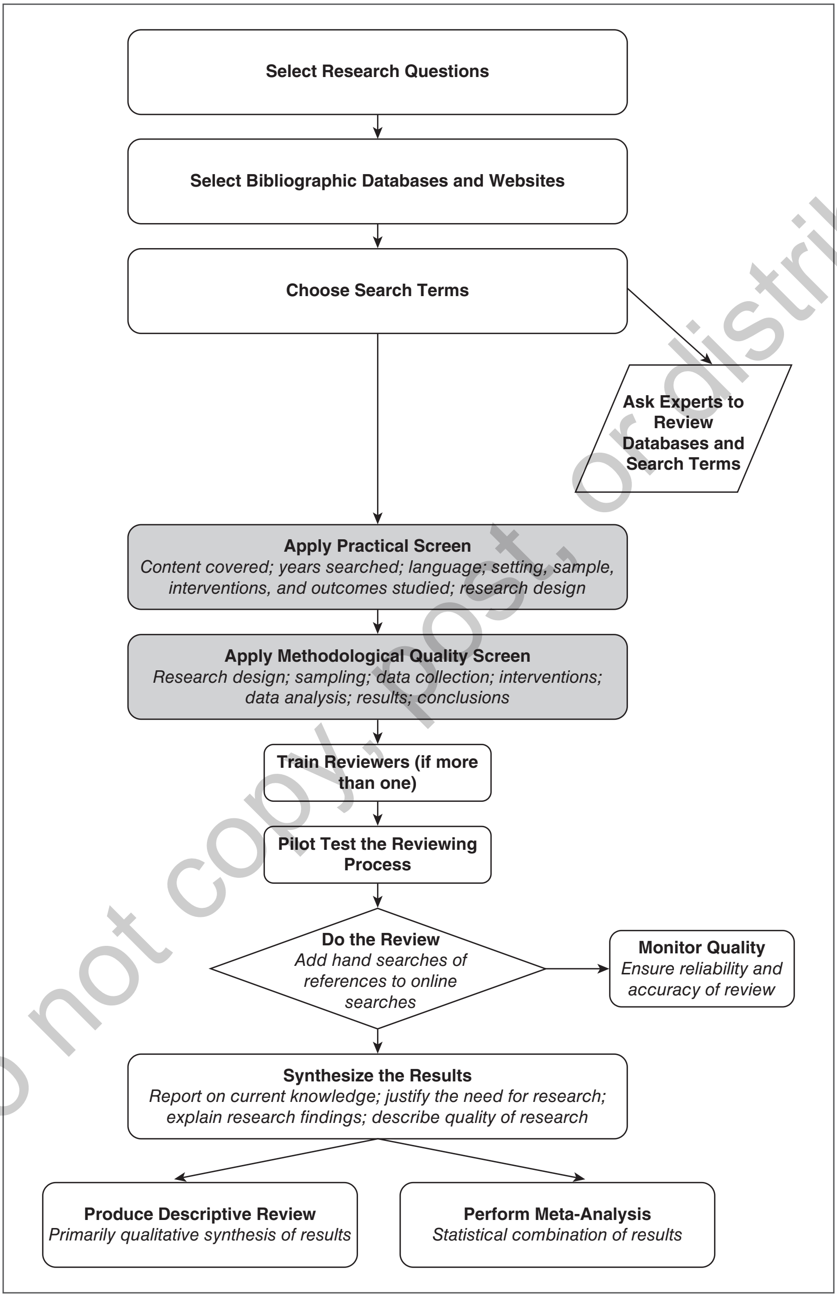
An overview of the basic components of research design and sampling—two components of methodological quality—is given in this chapter. The next chapter explains data collection, analysis, and reporting.

Figure 2.1 illustrates the steps in conducting research literature reviews. The shaded portions are covered in this chapter, applying the practical and methodological screens to the search.

A literature search that has no restrictions may yield hundreds of candidate articles for review. It is unlikely, however, that you will want to review all of them because many will be irrelevant or poorly designed. Some articles will be published in a language you cannot read, for example, and others might focus on topics that are not on target. If you are interested in reviewing articles on how to prevent the common cold, for example, a search online will produce articles on viruses that cause colds, the psychological effects of having a cold, methods of treatment, and so on. Some articles might be useful, but others will not. Before beginning to review them all, you must sort through them to identify the ones that contain information on prevention.

Suppose you find 50 studies that focus on your general topic: Preventing colds. Even then, you cannot assume that you have finished your search. In all likelihood, some studies will be methodologically rigorous, deriving sound conclusions from valid evidence, whereas others will be methodologically weak. To ensure the accuracy of your review, you must continue the screening process so that you can correctly distinguish well-designed studies from poorly designed ones.

Figure 2.1 Steps in Conducting Research Literature Reviews


Select Research Questions
Select Bibliographic Databases and Websites
Choose Search Terms
Ask Experts to Review Databases and Search Terms
Apply Practical Screen
Content covered; years searched; language; setting, sample, interventions, and outcomes studied; research design
Apply Methodological Quality Screen
Research design; sampling; data collection; interventions; data analysis; results; conclusions
Train Reviewers (if more than one)
Pilot Test the Reviewing Process
Do the Review
Add hand searches of references to online searches
Monitor Quality
Ensure reliability and accuracy of review
Synthesize the Results
Report on current knowledge; justify the need for research; explain research findings; describe quality of research
Produce Descriptive Review
Primarily qualitative synthesis of results
Perform Meta-Analysis
Statistical combination of results

Efficient searches use two screens to sort out the relevant and strong studies from the others. The first screen is primarily practical. It is used to identify a broad range of articles that may be potentially usable in that they cover the topic of interest, are in a language you read, and are in a publication you respect and can obtain in a timely manner. The second screen is for quality, and it helps you narrow your search by identifying the **best available** studies. The best studies are not trying to sell you anything and use the methods that scientists and scholars rely on to gather sound evidence. Screening articles for **methodological quality** is essential in ensuring the accuracy of your review.

You must use both search screens—**practical and methodological**—to ensure the review's efficiency, relevance, and accuracy.

Search Screen 1: The Practical Screen

The following are examples of the variety of practical screening criteria that might be used to guide your search.

Including and Excluding Studies: Typical Practical Screening Criteria for Literature Review Searches

1. **Publication language**

   Example. Include only studies in English and Spanish.

2. **Journal**

   Example. Include all education journals. Exclude all sociology journals.

3. **Author**

   Example. Include all articles by Wendy Adams.

4. **Setting**

   Example. Include all studies that take place in community health settings. Exclude all studies that take place in community social service centers.

5. **Participants or subjects**

   Example. Include all men and women. Include all people who have a valid driver's license. Exclude all people who will not take the driving test in English or Spanish.

6. **Program/interention**

   Example. Include all programs that are teacher led. Exclude all programs that are learner initiated.

7. **Research design**

   Example. Include only randomized trials/true experiments. Exclude studies without participant blinding.

8. **Sampling**

   Example. Include only studies that rely on randomly selected participants.

9. **Date of publication**

   Example. Include only studies published from January 1, 2005 to December 31, 2012.

10. **Date of data collection**

    Example. Include only studies that collected data from 2000 through 2012. Exclude studies that do not give dates of data collection.

11. **Duration of data collection**

    Example. Include only studies that collect data for 12 months or longer.

12. **Content (topics, variables)**

    Example. Include only studies that focus on primary prevention of illness. Exclude studies that focus on secondary or tertiary prevention. Exclude studies that focus on treatment.

13. **Source of financial support**

    Example. Include all privately supported studies. Exclude all studies receiving any government funds.

A literature review may use some or all types of practical screening criteria, as illustrated in these examples.

Practical Screening Criteria: Using Inclusion and Exclusion Criteria

Example 1. Social Functioning

To identify articles in English pertaining to measures of social functioning, we used three sources of information: The Oishi Social

Inclusion Criteria	Type
Term *social functioning* in titles	Content
Published from 2008 to the present	Publication date
Described or used at least one questionnaire	Content or instrument
English, French, Russian, Danish, or Spanish	Publication language
In 1 of 15 prominent journals (actual names given)	Journal
Exclusion Criteria	**Type**
Letters, editorials, review articles	Research design
Articles that deal with research design measure development or policy	Content

Functioning Bibliography (which cites 1,000 articles), PubMed (National Library of Medicine), and PsycINFO (American Psychological Association). We limited candidate articles to those having the term *social functioning* in their titles. From these candidate articles, we selected only those that were published from 2008 to the present and that also described or used at least one questionnaire. We excluded letters, editorials, reviews, and articles that either were not written in English, French, Russian, Danish, or Spanish or dealt primarily with methodology or policy. We then reviewed the list of articles and restricted our selection to 15 prominent journals. Here is a summary of the **inclusion and exclusion criteria**:

Example 2. Child Abuse and Neglect

We examined evaluations of programs to prevent child abuse and neglect conducted from 1990 through 2012. In our selection, we did not distinguish between types of abuse (such as physical or emotional) and neglect (such as emotional or medical), intensity, or frequency of occurrence. Only evaluations of programs that were family based, with program operations focused simultaneously on parents and children rather than just on parents, children, child care professionals, or the community, were included. We excluded studies that aimed to predict the causes and consequences of abuse or neglect or to appraise the effects of programs to treat children and families after abuse and neglect had been identified. We also excluded essays on abuse, cross-sectional studies, consensus statements, methodological research such as the development of a new measure of abuse, and studies that did not produce judgments of program effectiveness. Here is a summary of the inclusion and exclusion criteria:

Inclusion Criteria	Type
Evaluations of programs to prevent child abuse and neglect	Content
Conducted from 1990 through 2012	Duration of data collection
Family-based programs: focus on parents and families	Content
Exclusion Criteria	**Type**
Studies aiming to predict the causes and consequences of abuse or neglect	Content
Evaluations of programs to treat child abuse and neglect	Content
Essays on abuse, cross-sectional studies consensus statements, and studies that do not produce judgments of effectiveness	Research design
Methodological research such as the development of a new measure of abuse	Content

Search Screen 2: Methodological Quality Screening Criteria

The second screen—for methodological quality—consists of setting standards for high-quality studies. The idea is that you should review only the studies that meet the selected (and justified) standards. In practice, this means that your search will be considerably narrowed.

Methodological quality refers to how well—scientifically—a study has been designed and implemented to achieve its objectives. The highest quality studies come closest to adhering to rigorous research standards. Only methodologically sound studies produce accurate results. Focusing on sound studies is the only way to ensure the accuracy of the review. Because of this, learning about research methods is an essential component of a research literature review.

To select high-quality studies, the literature reviewer should ask the following: (a) Is this study's research design internally and externally valid? (b) Are the data sources used in the study reliable and valid? (c) Are the analytic methods appropriate given the characteristics and quality of the study's data? (d) Are the results meaningful in practical and statistical terms? As you will see from the discussion that follows, failure to provide satisfactory answers to some or all of these questions lessens a study's quality.

Criterion for Quality: Research Design

A study's **research design** refers to the way in which its subjects or participants—students, patients, and customers—are organized and measured. For instance, a study can be designed so that its participants are organized into two groups, one of which receives special treatment, but both are tested at least twice to find how they are.

Research designs are traditionally categorized as experimental or observational. In typical experimental studies, one or more groups of people participate in a new program or intervention (the "experiment"), and the researcher manipulates the environment to evaluate changes.

Experimental study designs typically involve two or more groups, at least one of which participates in an experiment while the other joins a **control (or comparison) group**, which does not take part in the experiment. The **experimental group** is given a new or untested, innovative program, intervention, or treatment. The control group is given an alternative. A group is any collective unit. Sometimes the unit is made up of individuals with a common experience, such as children who are in a reading program, people who fear heights, or scholarship winners. At other times, the unit is naturally occurring: a classroom, a business, or a hospital.

An example of an **experimental research design** might be one in which 100 teens are assigned at random to participate in a previously untried program to prevent high school dropout. The program includes work study classes and individual instruction to improve reading, writing, and computer skills. The progress of the students and their dropout rate are compared to another 100 teens who were randomly assigned to an alternative program. This program provides individual instruction but does not include work study.

Observational studies do not introduce new programs; they analyze already existing conditions and activities. An example of an observational study is one in which researchers analyze the school records of students to compare dropout rates between those who were in a prevention program and those who were not.

In general, experimental studies are considered more potent than observational designs. However, the use of experimental designs does not automatically guarantee a high-quality study, and it is important to learn about the characteristics of good research in order to make equitable and valid judgments.

Experimental Designs in Brief

Parallel Controls and Random Assignment: True Experiments. The groups in this design are created by first setting up eligibility criteria and then randomly assigning eligible *units* to one or more experimental and control groups. The groups can be observed and measured periodically. If the experimental group is observed to differ from the control group in a positive way on important variables (such as customer satisfaction, quality of life, health, and knowledge), the experiment is considered to be successful within certain predefined limits. The units that are randomly assigned may be individuals (such as Persons A, B, C, etc., or Teachers A, B, C, etc.) or clusters of individuals (such as schools, residential blocks, hospitals).

Random assignment (sometimes called **randomization** or **random allocation**) means that individuals or clusters of individuals are assigned by chance to the experimental or the control groups. With random assignment, the occurrence of previous events has no value in predicting future events. The alternative to randomization is regulation of the allocation process so that you can predict group assignment (such as assigning people admitted to a hospital on odd days of the month to the experimental group and those admitted on even days to the control group). A study with **parallel controls** and random assignment is illustrated next.

A Study With Parallel Controls and Random Assignment

Two reading programs were compared for students with reading difficulties. Over 4 years, half of all eligible children in each of 10 elementary schools were assigned at random to either Program A or Program B. Children were

Assigning Half of All Eligible Children in 10 Elementary Schools at Random to Program A or Program B

School	Intervention Groups	
	Program A	Program B
1 (100 children)	50	50
2 (60 children)	30	30
3 (120 children)	60	60
4 (90 children)	45	45
5 (100 children)	50	50
6 (90 children)	45	45
7 (70 children)	35	35
8 (150 children)	75	75
9 (150 children)	75	75
10 (100 children)	50	50

eligible if they were reading one or more grade levels below expectation as measured by the XYZ Reading Test. The design can be illustrated as follows:

Random selection is different from random assignment. In some studies, the entire eligible population is chosen. In others, a sample or fraction of the population is chosen. If this fraction is selected randomly, you have random selection. If you next decide to randomly assign this selected sample into two or more groups, you also have random assignment, as illustrated next.

Random Selection and Random Assignment: Two Examples

1. In Study A, teens who volunteered to participate in an evaluation of an experimental web-based history class were assigned at random to the experimental or a "control" program.
2. In Study B, a sample of teens was randomly selected from all who were eligible and then randomly assigned to the experimental or control program.

Comment. In the first study, teens were not selected at random but were chosen from a group of volunteers. Once chosen, however, they were assigned at random to the experimental or control program. In the second study, teens were randomly selected and randomly assigned. In general, random selection and random assignment are preferable to either alone.

Experimental study designs with randomly constituted parallel groups are the gold standards or the preferred designs when doing **scientific** research. They are sometimes referred to as *true experiments.* These designs—when implemented properly—can control for many errors or **biases** that may affect any experiment.

What are these errors or biases that lead to false conclusions? One of the most potentially damaging biases comes from the method of "selection." **Selection bias** is present when people who are initially different from one another and have differing prior risks for the outcome of interest are compared. Suppose a study is conducted after Schools 1 and 2 participate in a comparative test of two approaches to reading. The study results reveal that children in School 1's reading program—the control—score higher (better) on an attitude-to-reading inventory than do children in School 2—the experiment. Although the results may suggest a failed experiment, the two groups may have been different to begin with, even if they appeared to be similar. For instance, School 1's and 2's children may be alike in socioeconomic background, reading ability, and the competence of their reading teachers, but they may differ in other important ways. School 2, for example, may have a better library, a friendlier librarian, more resources to spend on extra program reading, a social system that reinforces reading, and so on. Less bias from the selection process would have resulted had students been randomly assigned into experimental and control groups regardless of school.

Biases can arise from unanticipated and unrecognized as well as recognized characteristics. Randomization is the only known way to control for unknown biases and to distribute them evenly among groups.

Designs using parallel controls and random assignment are more complex than other types of study designs. One issue that often arises in connection with these designs concerns the appropriate unit of randomization. Sometimes, for practical purposes, clusters (schools, companies), rather than individuals, are chosen for random assignment. When this happens, you cannot assume that the individuals forming the groups are comparable in the same way as they would have been had they been randomly chosen as individuals to be in the particular school or company. The reason is that if the researchers choose the assignment, they are in control, but if the individual participant or "subject" makes the selection, he or she is. After all, people go to certain schools, for example, because the schools meet the need of the individual and not of the experiment.

Other potential sources of bias include failure to adequately monitor the randomization process and to follow uniform procedures of randomization across all groups in the study. Training the people who do the randomizing and monitoring the quality of the process are essential, and the literature reviewer should be able to find these procedures discussed in the study.

In some randomized studies, the participants and investigators do not know which participants are in the experimental or the control groups: This is the double-blind experiment. When participants do not know, but investigators do, this is called the blinded trial. Some experts maintain that **blinding** is as important as randomization. Randomization, they say, eliminates influences at the start of a study but not the *confounders* that occur during the course of the study. For instance, confounding can occur if participants get extra attention or the control group catches on to the experiment. The extra attention and changes in the control group may alter the **outcomes** of a study. Blinding is often difficult to achieve in social experimentation, so the wary reviewer should pay special attention to the biases that may have occurred in randomized controlled studies without blinding.

Despite its scientific virtues, you cannot assume that randomization alone guarantees that a study has produced “truth.” At the minimum, valid study results also depend on accurate data collection and appropriate statistical analysis and interpretation.

For examples of randomized controlled trials, go to the following sources:

Baird, S. J., Garfein, R. S., McIntosh, C. T., & Ozler, B. (2012). Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: A cluster randomised trial. *Lancet, 379*(9823), 1320–1329. doi: 10.1016/s0140-6736(11)61709-1

Buller, M. K., Kane, I. L., Martin, R. C., Grese, A. J., Cutter, G. R., Saba, L. M., & Buller, D. B. (2008). Randomized trial evaluating computer-based sun safety education for children in elementary school. *Journal of Cancer Education, 23,* 74–79.

Butler, R. W., Copeland, D. R., Fairclough, D. L., Mulhern, R. K., Katz, E. R., Kazak, A. E., . . . Sahler, O. J. Z. (2008). A multicenter, randomized clinical trial of a cognitive remediation program for childhood survivors of a pediatric malignancy. *Journal of Consulting and Clinical Psychology, 76,* 367–378.

DuMont, K., Mitchell-Herzfeld, S., Greene, R., Lee, E., Lowenfels, A., Rodriguez, M., & Dorabawila, V. (2008). Healthy Families New York (HFNY) randomized trial: Effects on early child abuse and neglect. *Child Abuse & Neglect, 32,* 295–315.

Fagan, J. (2008). Randomized study of a prebirth coparenting intervention with adolescent and young fathers. *Family Relations, 57,* 309–323.

Gallardo, A., Hossain, S., Perez, E. A., Martin, J. Y., Lasalvia, S., Wong, K. A., . . . Sehgal, A. R. (2012). Effect of an iPod video intervention on consent to donate organs: A randomized trial. *Annals of Internal Medicine, 156*(7), 483–490. doi: 10.1059/0003-4819-156-7-201204030-00004

Johnson, J. E., Friedmann, P. D., Green, T. C., Harrington, M., & Taxman, F. S. (2011). Gender and treatment response in substance use treatment-mandated parolees. *Journal of Substance Abuse Treatment, 40*(3), 313–321. doi: 10.1016/j.jsat.2010.11.

Kubiak S. P., Kim, W. J., Fedock, G., & Bybee, D. (2015). Testing a violence-prevention intervention for incarcerated women using a randomized control trial. *Research on Social Work Practice*, *25*(3), 334–348.

Nance, D. C. (2012). Pains, joys, and secrets: Nurse-led group therapy for older adults with depression. *Issues in Mental Health Nursing, 33*(2), 89–95. doi: 10.3109/01612840.2011.624258

Poduska, J. M., Kellam, S. G., Wang, W., Brown, C. H., Ialongo, N. S., & Toyinbo, P. (2008). Impact of the Good Behavior Game, a universal classroom-based behavior intervention, on young adult service use for problems with emotions, behavior, or drugs or alcohol. *Drug and Alcohol Dependence, 95,* S29–S44.

Rdesinski, R. E., Melnick, A. L., Creach, E. D., Cozzens, J., & Carney, P. A. (2008). The costs of recruitment and retention of women from community-based programs into a randomized controlled contraceptive study. *Journal of Health Care for the Poor and Underserved, 19,* 639–651.

Swart, L., van Niekerk, A., Seedat, M., & Jordaan, E. (2008). Paraprofessional home visitation program to prevent childhood unintentional injuries in low-income communities: A cluster randomized controlled trial. *Injury Prevention, 14*(3), 164–169.

Thornton, J. D., Alejandro-Rodriguez, M., Leon, J. B., Albert, J. M., Baldeon, E. L., De Jesus, L. M., . . . Sehgal, A. R. (2012). Effect of an iPod video intervention on consent to donate organs: A randomized trial. *Annals of Internal Medicine, 156*(7), 483–490. doi: 10.1059/0003-4819-156-7-201204030-00004

Vaughn, S., Martinez, L. R., Wanzek, J., Roberts, G., Swanson, E., & Fall, A-M. (2017). Improving content knowledge and comprehension for English language learners: Findings from a randomized control trial. *Journal of Educational Psychology, 109*(1), 22–34.

Parallel Controls Without Random Assignment. Nonrandomized, parallel controls (**quasi-experimental designs** or nonequivalent control groups design) come about when you have at least two already existing groups, one of which is designated experimental. In education, the researcher might choose two comparable classrooms or schools and designate one as experimental. In community-based research, the researcher might use two similar communities. Researchers aim for groups that are as similar as possible so that they can be compared fairly or without bias. Unfortunately, the researchers can never be sure the groups are comparable. It is unlikely that the two groups would be as similar as they would if the researcher had assigned them through a random lottery. Here is an illustration.

Parallel Controls but No Random Assignment

A **nonrandomized** trial was used to test a program to reduce the use of antipsychotic drugs in nursing homes. The program was based on behavioral techniques to manage behavior problems and encourage gradual antipsychotic drug withdrawal. Two rural community nursing homes with elevated antipsychotic use were in the experimental group, and two other comparable homes were selected as parallel controls. Residents in both groups of homes had comparable demographic characteristics and functional status, and each group had a baseline rate of 29 days of antipsychotic use per 100 days of nursing home residence.

The above example uses a type of quasi-experimental design called a nonequivalent control groups design. Other quasi-experimental designs include the **time-series design** and its variations.

Parallel control designs without randomization are easier and less costly to implement than experimental designs with randomization, and many researchers use them. But these designs increase the likelihood that external **factors** will bias the results. Because of this, they are sometimes called *quasi-experimental* designs. A typical bias associated with nonrandom assignment is selection or membership bias.

Membership bias refers to the characteristics that members of groups share simply because they are in the group. The idea is that preexisting groups are usually not assembled haphazardly: They come together precisely because they share similar values, attitudes, behavior, or social and health status. Examples of groups with shared characteristics are people who live in the same neighborhood (who are likely to be similar in their incomes), children who have the same teacher (who may share similar abilities), patients who see a particular physician (who may have a particular medical problem), prisoners at a minimum-security facility (who have committed a certain level of crime), and prisoners at a maximum-security facility (who also have committed a certain level of crime and one that differs from those of prisoners in a minimum-security facility). Only random assignment can guarantee the limits within which two groups are equivalent from the point of view of all variables that may influence a study's outcomes.

Membership bias can seriously challenge a study's accuracy. When researchers use parallel controls without random assignment, you should look to see if they have administered a premeasure to determine the equivalence of the groups on potentially important characteristics at the study's start. In the study described above, the researchers demonstrate the equivalence of the groups by reporting that residents in each of the two homes had comparable demographic characteristics, functional status, and use of antipsychotics.

Statistical methods (e.g., propensity, instrumental variables methods, and regression discontinuity) are available to "control" for the influence of **confounding variables** (also referred to as **mediator variables**) when random assignment is not used. A variable that is more likely to be present in one group of subjects than in the comparison group and that is related to the outcome of interest and confuses or confounds the results is called a confounding variable. As a rule, however, it is better to control for confounders before the researchers collect data—that is, as part of design and sampling—than afterward, during analysis. You should review the design, sampling, and statistical analysis sections of the article to find out if the researchers adequately dealt with confounding.

For examples of nonrandomized controlled trails or quasi-experimental studies, go to the following sources:

Corcoran, J. (2006). A comparison group study of solution-focused therapy versus "treatment-as-usual" for behavior problems in children. *Journal of Social Service Research, 33,* 69–81.

Cross, T. P., Jones, L. M., Walsh, W. A., Simone, M., & Kolko, D. (2007). Child forensic interviewing in Children's Advocacy Centers: Empirical data on a practice model. *Child Abuse & Neglect, 31,* 1031–1052.

Gatto, N. M., Ventura, E. E., Cook, L. T., Gyllenhammer, L. E., & Davis, J. N. (2012). LA Sprouts: A garden-based nutrition intervention pilot program influences motivation and preferences for fruits and vegetables in Latino youth [Article]. *Journal of the Academy of Nutrition and Dietetics, 112*(6), 913–920. doi: 10.1016/j.jand.2012.01.014

Hebert, R., Raiche, M., Dubois, M. F., Gueye, N. R., Dobuc, N., Tousignant, M., & The PRISMA Group. (2010). Impact of PRISMA, a coordination-type integrated service delivery system for frail older people in Quebec (Canada): A quasi-experimental study [Article]. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences, 65*(1), 107–118. doi: 10.1093/geronb/gbp027

Hooshmand, M., & Foronda, C. (2017). Comparison of telemedicine to traditional face-to-face care for children with special needs: A quasi-experimental study. *Telemedicine and e-Health, 24*(6).

Jackson, N. J., Isen, J. D., Khoddam, R., Irons, D., Tuvblad, C., Iacono, W. G., . . . Baker, L. A. (2016). Impact of adolescent marijuana use on intelligence: Results from two longitudinal twin studies. *Proceedings of the National Academy of Sciences, 113*(5), E500–E508.

Jadallah, M., Hund, A. M., Thayn, J., Studebaker, J. G., Roman, Z. J., & Kirby, E. (2017). Integrating geospatial technologies in fifth-grade curriculum: Impact on spatial ability and map-analysis skills. *Journal of Geography*, *116*(4), 139–151.

Kutnick, P., Ota, C., & Berdondini, L. (2008). Improving the effects of group working in classrooms with young school-aged children: Facilitating attainment, interaction and classroom activity. *Learning and Instruction, 18,* 83–95.

Meizlish, D. S., Wright, M. C., Howard, J., & Kaplan, M. L. (2018). Measuring the impact of a new faculty program using institutional data. *International Journal for Academic Development, 23*(2), 72–85.

Orthner, D. K., Cook, P., Sabah, Y., & Rosenfeld, J. (2006). Organizational learning: A cross-national pilot-test of effectiveness in children's services. *Evaluation and Program Planning, 29,* 70–78.

Pascual-Leone, A., Bierman, R., Arnold, R., & Stasiak, E. (2011). Emotion-focused therapy for incarcerated offenders of intimate partner violence: A 3-year outcome using a new whole-sample matching method [Article]. *Psychotherapy Research, 21*(3), 331–347. doi: 10.1080/10503307.2011.572092

Reiser, J. E., & McCarthy, C. J. (2018). Preliminary investigation of a stress prevention and mindfulness group for teachers. *The Journal for Specialists in Group Work, 43*(1), 2–34.

Rice, V. H., Weglicki, L. S., Templin, T., Jamil, H., & Hammad, A. (2010). Intervention effects on tobacco use in Arab and non-Arab American adolescents [Article]. *Addictive Behaviors, 35*(1), 46–48. doi: 10.1016/j.addbeh.2009.07.005

Struyven, K., Dochy, F., & Janssens, S. (2008). The effects of hands-on experience on students' preferences for assessment methods. *Journal of Teacher Education, 59,* 69–88.

VanGarde, A., Yoon, J., Luck, J., & Mendez-Luck, C. A. (2018). Racial/ethnic variation in the impact of the Affordable Care Act on insurance coverage and access among young adults. *American Journal of Public Health*, *108*(4), 544–549.

Self-Controls

A design with self-controls uses a group of participants to serve as its own comparison. Suppose, for example, students were surveyed three times: at the beginning of the year to find out their attitudes toward community service, immediately after their participation in a 1-year course to find out the extent to which their attitude changed, and at the end of 2 years to ascertain if the change is sustained. This three-measurement strategy describes a design using the students as their own control. In the example, the survey measures the students once before and twice after the intervention (a new course).

Self-controlled designs are extremely weak because they are prone to many biases. Participants may become excited about taking part in an experiment; they may mature physically, emotionally, and intellectually; or historical events can intervene. For example, suppose a study reveals that the students in a 2-year test of a school-based intervention acquire important attitudes and behaviors and retain them over time. This desirable result may be due to the new course or to the characteristics of the students who, from the start, may have been motivated to learn and have become even more excited by being in an experimental program. Another possibility is that over the 2-year intervention period, students may have matured intellectually, and this development rather than the program is responsible for the learning. Also, historical or external events may have occurred to cloud the effects of the new course. For example, suppose that during the year, an inspired teacher gives several stimulating lectures to the students. The students' outstanding performance on subsequent tests may be due as much or more to the lectures as to the program.

The addition of a control group is necessary to strengthen self-controlled designs, as shown next.

Combined Self-Control and Parallel Control Design to Evaluate the Impact of Education and Legislation on Children's Use of Bicycle Helmets

An anonymous questionnaire regarding use of bicycle helmets was sent twice to nearly 3,000 children in three counties. The first mailing took place 3 months before an educational campaign in County 1 and 3 months before

Percentage of Children Reporting Helmet Use "Always" or "Usually"

	Before Intervention	After Intervention
County 1: Education only	8	13[a]
County 2: Education and legislation	11	37[b]
County 3: No intervention	7	8

Note: The percentages are small and do not add up to 100% because they represent just the proportion of children answering "always" or "usually." Other responses (such as "rarely") constituted the other choices.

a. $p < .01$. This means that the observed result (always or usually reporting helmet use) will occur by chance 1 in 100 times. The *p* or *p* value is the probability that an observed result (or result of a statistical test) is due to chance.

b. $p < .0001$. This means that the observed result will occur by chance 1 in 10,000 times.

the passage of legislation requiring helmets and an education campaign in County 2. The second mailing took place 9 months after completion of the education and combined education-legislation. Two surveys (9 months apart) were also conducted in County 3, the control. County 3 had neither education nor legislation pertaining to the use of bicycle helmets. The table and associated text summarize the results.

Note. More information about *p* values and other statistical terms can be found in Chapter 3.

Findings. The proportion of children who reported that they "always" or "usually" wore a helmet increased significantly ($p < .0001$) from 11% before to 37% in County 2 (education and legislation) and 8% to 13% ($p < .01$) in County 1 (education only). The increase of 1% in County 3 was not **statistically significant**.

Comment. Education alone and education combined with legislation were relatively effective: Either one or both increase the proportion of children reporting helmet use. The education may have taught children to give the socially acceptable responses on the survey, but other studies in the literature suggest that single education programs alone have not usually encouraged children to give desirable responses to survey questions. The fact that the control group did not improve suggests that County 1's and County 2's efforts were responsible for the improvements. The addition of the control group adds credibility to the study results.

Historical Controls or Existing Data

Studies that use **historical controls** rely on data that are available from an existing database. These data are sometimes called "norms" to refer to the fact they are the reference points. They are historical because they were collected in the past.

Historical controls include established norms such as scores on standardized tests (e.g., the SATs and GREs), the results of studies conducted in the past, and vital statistics such as birth and death rates. These data substitute for the data that would come from a parallel control.

Suppose you are reviewing the literature to find out how your state or province compares with the rest of the country in its provision of routine

Table 2.1 Percentages of School Children in a Hypothetical State (HS) and Country (HC) Who Have Access to Routine School-Based Mental Health Services

	≤10 Years (%)		10 to 14 Years (%)		15 to 17 Years (%)	
	HS	HC	HS	HC	HS	HC
All children	89.9	97.2	92.3	84.5	89.5	90.8
Family income						
Under $25,000	83.7	82.4	86.7	86.3	87.6	88.8
$25,000 to $50,000	95.7	92.7	91.3	92.5	86.5	88.3
$51,000 to $75,000	95.3	91.3	96.1	94.6	90.1	89.4
Greater than $75,000	95.3	94.4	97.7	96.5	96.7	94.7

Sources: National Survey of Children's Mental Services (2003) and State Survey of Children's Mental Health Services (1998).

school-based mental health services. Your state has just completed a survey of its schools. You come across Table 2.1 in a report.

The people who prepared the table have used historical controls in the form of an existing database (National Survey of Children's Mental Health Services, 2003) as a reference against which to compare the results of the more recent survey. (The statistical methods for comparing the results are not included in this example.)

Historical controls are convenient; their main source of bias is the potential lack of comparability between the group on whom the data were originally collected and the group of concern in the study. In the example given in the table, the reviewer has to determine if children in the state have different needs or resources compared with the rest of the county. If so, then the comparison group is not appropriate.

The validity of the comparisons between historical controls and current groups may also be compromised if the data come from two different time periods because of rapid social changes. In the example in Table 2.1, the reviewer might fault the researcher for using data that are 10 (or more) years old, particularly if evidence shows that mental health services, needs, and resources have changed markedly from over the data collection period.

When reviewing the literature, ask the following: Is the choice of historical control explained and justified? Are the normative data reliable, valid, and appropriate?

Observational Designs in Brief

Cohort Designs

A **cohort** is a group of people who have something in common and who remain part of a study group over an extended period of time. In public health research, cohort studies are used to describe and predict the risk factors for a disease and the disease's cause, incidence, natural history, and prognosis. Cohort studies may be **prospective** or **retrospective**. With a prospective design, the direction of inquiry is forward in time, whereas with a retrospective design, the direction is backward in time.

Prospective Cohort Designs

Does a High-Fiber Diet Prevent Colon Cancer?

A team of researchers was interested in finding out if a high-fiber diet prevents colon cancer. They mailed out questionnaires to a sample of registered nurses (the cohort) asking them about their diet and other risk factors and received over 121,000 completed responses. Every 2 years for 2 decades, they sent out questionnaires to update their information and to ask the nurses about the occurrence of any diseases, including colon cancer. The researchers confirmed the nurses' report of disease by examining their medical records. The statistical analysis of the data showed that dietary fiber intake did not prevent colon cancer. Nurses who consumed the least amount of dietary fiber did not differ from nurses who consumed the most in rates of colon cancer.

This is a very brief description of one small component of the Nurses' Health Study, a large multiyear cohort study. With rigorous cohort designs, potential *predictive* factors (such as diet) are measured before an *outcome* (such as colon cancer) occurs. Over a long time period and with multiple and frequent valid measures, the researcher may be able to infer that the factor is (or is not) a cause of the outcome.

Another example of a prospective cohort study involves the study of criminal careers.

(Continued)

(Continued)

Do Criminal Career Patterns Differ Across Race and by Sex?

Researchers analyzed data from individuals (the cohort) who participated in the Providence sample of the National Collaborative Perinatal Project. They focused on patterns of prevalence, frequency, chronicity, and specialization in violence for the entire cohort, as well as for samples stratified by race, sex, and race together with sex. In addition, demographic and juvenile offending characteristics were used to predict adult offender status. The researchers found that three variables significantly predicted adult offender status. Males and non-Whites were significantly more likely than females and Whites to be registered as adult offenders. Of the two juvenile offending indicators, only one, chronic offending, significantly predicted adult offender status. Having a violent arrest as a juvenile did not significantly predict adult offender status.

Question: Which were the predictor(s) in this study? Which were the outcomes?

High-quality prospective or **longitudinal studies** are expensive to conduct, especially if the investigator is concerned with outcomes that are relatively rare or hard to predict. Studying rare and unpredictable outcomes requires large samples and numerous measures. Also, researchers who do prospective cohort studies have to be on guard against loss of subjects over time or **attrition**. For instance, longitudinal studies of children are often beset by attrition because over time, they lose interest, move far away, change their names, and so on. If a large number of people drop out of a study, the sample that remains may be very different from the one that left. The remaining sample may be more motivated or less mobile than those who left, for example, and these factors may be related in unpredictable ways to any observed outcomes.

When reviewing prospective cohort studies, make sure that the researchers address how they handled **loss to follow-up** or attrition. Ask the following: How large a problem was attrition? Were losses to follow-up handled in the analysis? Were the study's findings affected by the losses?

For examples of prospective cohort studies, go to the following:

Brown, C. S., & Lloyd, K. (2008). OPRISK: A structured checklist assessing security needs for mentally disordered offenders referred to high security psychiatric hospital. *Criminal Behaviour and Mental Health, 18,* 190–202.

Chauhan, P., & Widom, C. S. (2012). Childhood maltreatment and illicit drug use in middle adulthood: The role of neighborhood characteristics. *Development and Psychopathology, 24*(Special Issue 03), 723–738. doi: 10.1017/S0954579412000338

Fuchs, C. S., Giovannucci, E. L., Colditz, G. A., Hunter, D. J., Stampfer, M. J., Rosner, B., . . . Willett, W. C. (1999). Dietary fiber and the risk of colorectal cancer and adenoma in women. *New England Journal of Medicine, 340,* 169–176.

Kemp, P. A., Neale, J., & Robertson, M. (2006). Homelessness among problem drug users: Prevalence, risk factors and trigger events. *Health & Social Care in the Community, 14,* 319–328.

Kerr, T., Hogg, R. S., Yip, B., Tyndall, M. W., Montaner, J., & Wood, E. (2008). Validity of self-reported adherence among injection drug users. *Journal of the International Association of Physicians in AIDS Care, 7*(4), 157–159.

Kuyper, L. M., Palepu, A., Kerr, T., Li, K., Miller, C. L., Spittal, p. m., . . . Wood, E. (2005). Factors associated with sex-trade involvement among female injection drug users in a Canadian setting. *Addiction Research & Theory, 13*(2), 193–199.

Piquero, A. R., & Buka, S. L. (2002). Linking juvenile and adult patterns of criminal activity in the Providence cohort of the National Collaborative Perinatal Project. *Journal of Criminal Justice, 30,* 259–272.

Pletcher, M. J., Vittinghoff, E., Kalhan, R., Richman, J., Safford, M., Sidney, S., Lin, F., & Kertesz, S. (2012). Association between marijuana exposure and pulmonary function over 20 years. *JAMA: The Journal of the American Medical Association, 307*(2), 173–181. doi: 10.1001/jama.2011.1961

White, H. R., & Widom, C. S. (2003). Does childhood victimization increase the risk of early death? A 25-year prospective study. *Child Abuse & Neglect, 27,* 841–853.

Because of the difficulties and expense of implementing prospective cohort designs, many cohort designs reported in the literature tend to be retrospective. An example retrospective cohort design is illustrated as follows.

Retrospective Cohort Design

Does a Relationship Exist Between Family History and Diagnosis of Breast Cancer?

The investigators searched a hospital's database of diagnoses between 2000 and 2004 and found 250 women who had a diagnosis of cancer in situ. These 250 women are the cohort. The investigators reviewed these 250 patients' medical records to find out about their family's medical history and about other factors that might be associated with the disease. The data collected by

(*Continued*)

(Continued)

the researchers enabled them to study the relationship between family history and other variables and the occurrence of cancer in this sample of 250.

What Are the Possible Causes of Xenophobia in German Youth?

Using available data from a large, ongoing study of youths from East and West Berlin, trends of change in adolescent xenophobia were analyzed. The study's database contained the results of two surveys: the Self-Interested Survey and the Self-Esteem Scale. Two main hypotheses were tested—namely, that self-interest and low self-esteem are the driving forces behind xenophobia among 13- to 16-year-olds.

Retrospective cohort designs have the same strengths as prospective designs. Like prospective designs, retrospective designs can establish that a predictor variable (such as self-esteem) precedes an outcome (such as xenophobia). Also, because data are collected before the outcomes are known, the measurement of variables that might predict the outcome (such as self-esteem) cannot be biased by prior knowledge of which people are likely to develop the problem (such as xenophobia). Retrospective cohort studies are usually less expensive to do than prospective studies because they rely on existing data. But the results may not be as convincing because the existing data on which the investigator depends may not include the subjects and information that the investigator might prefer if he or she had done the original study. When you review a retrospective cohort study, ask the following: How typical or representative is the sample? Is the investigation of the cohort inclusive? Did the analysis include all pertinent variables?

For an example of retrospective cohort studies, go to the following:

Boehnke, K., Hagan, J., & Hefler, G. (1998). On the development of xenophobia in Germany: The adolescent years. *Journal of Social Issues* [Special issue: Political development: Youth growing up in a global community], *54,* 585–602.

Edith, H. H., Pierik, F. H., de Kluizenaar, Y., Willemsen, S. P., Hofman, A., van Ratingen, S. W., . . . Jaddoe, V. W. V. (2012). Air pollution exposure during pregnancy, ultrasound measures of fetal growth, and adverse birth outcomes: A prospective cohort study. *Environmental Health Perspectives, 120*(1), 150–156. doi: 10.1289/ehp.1003316

Harper, S., Rushani, D., & Kaufman, J. S. (2012). Trends in the black-white life expectancy gap, 2003-2008. *JAMA: The Journal of the American Medical Association, 307*(21), 2257–2259. doi: 10.1001/jama.2012.5059

Hoge, C. W., Auchterlonie, J. L, & Milliken, C. S. (2006). Mental health problems, use of mental health services, and attrition from military service after returning

from deployment to Iraq or Afghanistan. *JAMA: The Journal of the American Medical Association, 295*(9), 1023–1032. doi: 10.1001/jama.295.9.1023

Kerr, K., Romaniuk, M., McLeay, S., Khoo, A., Dent, M. T., & Boshen, M. (2018). Increased risk of attempted suicide in Australian veterans is associated with total and permanent incapacitation, unemployment and posttraumatic stress disorder severity. *Australian & New Zealand Journal of Psychiatry, 52*(6), 552–560.

Lee, M., Krishnamurthy, J., Susi, A., Sullivan, C., Gorman, G. H., Hisle-Gorman, E., . . . Nylund, C. M. (2018). Association of Autism Spectrum Disorders and inflammatory bowel disease. *Journal of Autism and Developmental Disorders*, *48*(5), 1523–1529.

Lee, S. J., Taylor, C. A., & Bellamy, J. L. (2012). Paternal depression and risk for child neglect in father-involved families of young children [Article]. *Child Abuse & Neglect, 36*(5), 461–469. doi: 10.1016/j.chiabu.2012.04.002

Lurie, I., Gur, A., Haklai, Z., & Goldberger, N. (2018). Suicide risk Among Holocaust survivors following psychiatric hospitalizations: A historic cohort study. *Archives of Suicide Research, 22*(3), 496–509.

Santos, I. S., Matijasevich, A., & Domingues, M. R. (2012). Maternal caffeine consumption and infant nighttime waking: Prospective cohort study. *Pediatrics, 129*(5), 860–868. doi: 10.1542/peds.2011–1773

Case Control Designs

Case control designs are generally retrospective. They are used to explain why a phenomenon currently exists by comparing the histories of two different groups, one of which is involved in the phenomenon. For example, a case control design might be used to help understand the social, demographic, and attitudinal variables that distinguish people who at the present time have been identified with frequent headaches from those who do not have frequent headaches.

The cases in case control designs are individuals who have been chosen on the basis of some characteristic or outcome (such as frequent headaches). The controls are individuals without the characteristic or outcome. The histories of cases and controls are analyzed and compared in an attempt to uncover one or more characteristics present in the cases and not in the controls.

How can researchers avoid having one group decidedly different from the other, say, healthier or smarter? Some methods include randomly selecting the controls, using several controls, and carefully **matching** controls and cases on important variables.

The Case Control Design in Two Studies

The Role of Alcohol in Boating Deaths

Alcohol is increasingly recognized as a factor in many boating fatalities, but the association between alcohol consumption and mortality among boaters has not been well quantified. This study aimed to determine the association of alcohol use with passengers' and operators' estimated relative risk of dying while boating. To do this, the researchers carried out a case control study of recreational boating deaths among persons aged 18 years or older from 1990 to 1998 in Maryland and North Carolina ($n = 221$). They compared the cases with control interviews obtained from a multistage probability sample of boaters from the same locations at which the deaths occurred in each state from 1997 to 1999 ($n = 3{,}943$).

Knee Arthritis and Japanese Women

This study investigated the relationship between knee osteoarthritis (OA) and constitutional factors (e.g., weight), history of joint injuries, and occupational factors using a case control study among women in Japan.

The study covered three health districts in Japan. Cases were women 45 years of age and older who were diagnosed with knee OA by orthopedic physicians using radiography. Controls were selected randomly from the general population and were individually matched to each case for age, sex, and residential district. Subjects were interviewed using structured questionnaires to determine medical history, including history of joint injury, physical activity, socioeconomic factors, and occupation. Height and weight were measured.

In the first study, a complex random sampling scheme was employed to minimize bias among control subjects and maximize their comparability with cases (e.g., deaths took place in the same location). In the second study, the controls were selected randomly and then were matched to each case for age, sex, and residential districts.

Epidemiologists and other health workers often use case control designs to provide insight into the causes and consequences of disease and other health problems. Reviewers of these studies should be on the lookout for certain methodological problems, however. First, cases and controls are often chosen from two separate populations. Because of this, systematic differences (such as motivation and cultural beliefs) may exist between or among the groups that are difficult to anticipate, measure, or control, and these differences may influence the study's results.

Another potential problem with case control designs is that the data often come from people's recall of events, such as asking women to discuss the history of their physical activity or asking boaters about their drinking habits. Memory is often unreliable, and if so, the results of a study that depends on memory may result in misleading information.

For examples of case control studies, go to the following:

Barney, C. C., Tervo, R., Wilcox, G. L., & Symons, F. J. (2017). A case-controlled investigation of tactile reactivity in young children with and without global developmental delay. *American Journal on Intellectual and Developmental Disabilities, 122*(5), 409–421. doi: 10.1352/1944-7558-122.5.409

Belardinelli, C., Hatch, J. P., Olvera, R. L., Fonseca, M., Caetano, S. C., Nicoletti, M., Pliszka, S., & Soares, J. C. (2008). Family environment patterns in families with bipolar children. *Journal of Affective Disorders, 107*(1–3), 299–305.

Bookle, M., & Webber, M. (2011). Ethnicity and access to an inner city home treatment service: A case-control study. *Health & Social Care in the Community, 19*(3), 280–288. doi: 10.1111/j.1365-2524.2010.00980.x

Davis, C., Levitan, R. D., Carter, J., Kaplan, A. S., Reid, C., Curtis, C., Patte, K., & Kennedy, J. L. (2008). Personality and eating behaviors: A case-control study of binge eating disorder. *International Journal of Eating Disorders, 41,* 243–250.

Grubbs, J. B., Wilt, J. A., Exline, J. J., & Pargament, K. I. (2018). Predicting pornography use over time: Does self-reported "addiction" matter? *Addictive Behaviors, 82*, 57–64. doi: https://doi.org/10.1016/j.addbeh.2018.02.028

Hall, S. S., Arron, K., Sloneem, J., & Oliver, C. (2008). Health and sleep problems in cornelia de lange syndrome: A case control study. *Journal of Intellectual Disability Research, 52,* 458–468.

Hourani, L. L., Williams, J., Lattimore, P. K., Morgan, J. K., Hopkinson, S. G., Jenkins, L., & Cartwright, J. (2018). Workplace victimization risk and protective factors for suicidal behavior among active duty military personnel. *Journal of Affective Disorders, 236*, 45–51. doi: https://doi.org/10.1016/j.jad.2018.04.095

Levi-Belz, Y., Gvion, Y., Grisaru, S., & Apter, A. (2018). When the pain becomes unbearable: Case-control study of mental pain characteristics among medically serious suicide attempters. *Archives of Suicide Research, 22*(3), 380–393. doi: 10.1080/13811118.2017.1355288

Marí-Bauset, S., Llopis-González, A., Zazpe-García, I., Marí-Sanchis, A., & Morales-Suárez-Varela, M. (2015). Nutritional status of children with Autism Spectrum Disorders (ASDs): A case–control study. [journal article]. *Journal of Autism and Developmental Disorders, 45*(1), 203–212.

(Continued)

(Continued)

Menendez, C. C., Nachreiner, N. M., Gerberich, S. G., Ryan, A. D., Erkal, S., McGovern, P. M., . . . Feda, D. M. (2012). Risk of physical assault against school educators with histories of occupational and other violence: A case-control study. *Work, 42*(1), 39–46.

Smith, G. S., Keyl, P. M., Hadley, J. A., Bartley, C. L., Foss, R. D., Tolbert, W. G., & McKnight, J. (2001). Drinking and recreational boating fatalities: A population-based case-control study. *Journal of the American Medical Association, 286,* 2974–2980.

Yoshimura, N., Nishioka, S., Kinoshita, H., Hori, N., Nishioka, T., Ryujin, M., . . . Cooper, C. (2004). Risk factors for knee osteoarthritis in Japanese women: Heavy weight, previous joint injuries, and occupational activities. *Journal of Rheumatology, 31*(1), 157–162.

Zerbo, O., Qian, Y., Yoshida, C., Grether, J. K., Van de Water, J., & Croen, L. A. (2015). Maternal infection during pregnancy and Autism Spectrum Disorders. [journal article]. *Journal of Autism and Developmental Disorders, 45*(12), 4015–4025. doi: 10.1007/s10803-013-2016-3

A Note on Other Designs and Studies: Cross-Sectional Surveys and Consensus Statements

Cross-Sectional Surveys

Cross-sectional designs result in a portrait of one or many groups at one period of time. These designs are frequently associated with mail and other **self-administered survey questionnaires** and face-to-face and telephone interviews. In fact, **cross-sectional studies** are sometimes called survey or descriptive designs. The following are three illustrative uses of cross-sectional designs.

Cross-Sectional Designs

1. Refugees are interviewed to find out their immediate fears and aspirations.
2. A survey is mailed to consumers to identify perceptions of the quality of the goods and services received when ordering by catalogue.
3. A community participates in a web survey to find out its needs for youth services.

Cross-sectional surveys are used to describe a study's sample and provide baseline information at the start of an experiment. The study's sample may consist of individuals or institutions such as businesses, schools, and hospitals. A researcher who conducts a web survey with 500 small businesses to find out their maternity leave policies is doing a cross-sectional study.

Baseline information consists of demographic data (age, gender, income, education, health) and statistics on variables such as current knowledge, attitudes, and behaviors. A researcher may, however, look for relationships between demographic data and other variables. For instance, a cross-sectional survey of knowledge of current events among middle school children might study the relationship between gender (one baseline variable) and knowledge of current events (another baseline variable).

The major limitation of cross-sectional studies is that on their own and without follow-up, they provide no information on causality: They only provide information on events at a single, fixed point in time. For example, suppose a researcher finds that girls have less knowledge of current events than do boys. The researcher cannot conclude that being female somehow causes less knowledge of current events. The researcher can only be sure that in *this* survey, girls had less knowledge than boys.

To illustrate the point further, suppose you are doing a literature review on community-based exercise programs. You are specifically interested in learning about the relationship between age and exercise. Does exercise decrease with age? In your search of the literature, you find the following report.

A Report of a Cross-Sectional Survey of Exercise Habits

In March of this year, Researcher A surveyed a sample of 1,500 people between the ages of 30 and 70 to find out about their exercise habits. One of the questions he asked participants was, "How much do you exercise on a typical day?" Researcher A divided his sample into two groups: people 45 years of age and younger and people 46 years and older. Researcher A's data analysis revealed that the amount of daily exercise reported by the two groups differed, with the younger group reporting 15 minutes more exercise on a typical day.

(Continued)

(Continued)

Based on this summary, does amount of exercise decline with age? The answer is that you cannot get the answer from Researcher A's report. The decline seen in a cross-sectional study like this one can actually represent a decline in exercise with increasing age, or it may reflect the oddities of this particular sample. The younger people in this study may be especially sports minded, whereas the older people may be particularly antiexercise. As a reviewer, you need to figure out which of the two explanations is better. One way you can do this is to search the literature to find out which conclusions are supported by other studies. Does the literature generally sustain the idea that amount of exercise always declines with age? After all, in some communities, the amount of exercise done by older people may actually increase because with retirement or part-time work, older adults may have more time to exercise than do younger people.

Suppose you are interested in finding out how parents of children with Tourette's disorder and parents of children with asthma compare in their mental health and burden of caregiving. Can you get the information you need from the following cross-sectional survey?

A Cross-Sectional Survey of Two Groups of Parents

Researchers examined the mental health and caregiver burden in parents of children with Tourette's disorder compared with parents of children with asthma. They surveyed parents at Tourette's disorder and pediatric asthma hospital outpatient clinics. The survey consisted of measures of parent mental health (General Health Questionnaire [GHQ]-28) and caregiver burden (Child and Adolescent Impact Assessment). Of the parents of children with Tourette's, 76.9% had mental health distress on the GHQ-28, compared with 34.6% of the parents of children with asthma; this effect remained significant after taking into account demographic variables (such as age and education). Parents of children with Tourette's also experienced greater caregiver burden.

It is difficult to tell from the example if the differences in mental health found by the researchers are due to the nature of Tourette's parents' caregiving burden or to something else entirely. It is possible, for example, that the particular group of Tourette's parents might have significant mental health problems regardless of their children's illness. The reviewer needs more information about the two study samples and how they were selected in order to make a decision as to the validity of the researchers' findings.

For examples of cross-sectional studies, go to the following:

Belardinelli, C., Hatch, J. P., Olvera, R. L., Fonseca, M., Caetano, S. C., Nicoletti, M., Pliszka, S., & Soares, J. C. (2008). Family environment patterns in families with bipolar children. *Journal of Affective Disorders, 107*(1–3), 299–305.

Carmona, C. G. H., Barros, R. S., Tobar, J. R., Canobra, V. H., & Montequín, E. A. (2008). Family functioning of out-of-treatment cocaine base paste and cocaine hydrochloride users. *Addictive Behaviors, 33,* 866–879.

Cooper, C., Robertson, M. M., & Livingston, G. (2003). Psychological morbidity and caregiver burden in parents of children with Tourette's disorder and psychiatric comorbidity. *Journal of the American Academy of Child & Adolescent Psychiatry, 42,* 1370–1375.

Davis, C., Levitan, R. D., Carter, J., Kaplan, A. S., Reid, C., Curtis, C., Patte, K., & Kennedy, J. L. (2008). Personality and eating behaviors: A case-control study of binge eating disorder. *International Journal of Eating Disorders, 41,* 243–250.

Friedrichs, A., Silkens, A., Reimer, J., Kraus, L., Scherbaum, N., Piontek, D., . . . Buchholz, A. (2018). Role preferences of patients with alcohol use disorders. *Addictive Behaviors, 84*, 248–254. doi: https://doi.org/10.1016/j.addbeh.2018.05.002

Gorzycki, M., Howard, P., Allen, D., Desa, G., & Rosegard, E. (2016). An exploration of academic reading proficiency at the university level: A cross-sectional study of 848 undergraduates. *Literacy Research and Instruction, 55*(2), 142–162. doi: 10.1080/19388071.2015.1133738

Hall, S. S., Arron, K., Sloneem, J., & Oliver, C. (2008). Health and sleep problems in cornelia de lange syndrome: A case control study. *Journal of Intellectual Disability Research, 52,* 458–468.

Joice, S., Jones, M., & Johnston, M. (2012). Stress of caring and nurses' beliefs in the stroke rehabilitation environment: A cross-sectional study. *International Journal of Therapy & Rehabilitation, 19*(4), 209–216.

Kasbia, G. S., Farragher, J., Kim, S. J., Famure, O., & Jassal, S. V. (2014). A cross-sectional study examining the functional independence of elderly individuals with a functioning kidney transplant. *Transplantation, 98*(8), 864–870. doi: 10.1097/tp.0000000000000126

Kypri, K., Bell, M. L., Hay, G. C., & Baxter, J. (2008). Alcohol outlet density and university student drinking: A national study. *Addiction, 103,* 1131–1138.

Meijer, J. H., Dekker, N., Koeter, M. W., Quee, P. J., Van Beveren, N. J., & Meijer, C. J. (2012). Cannabis and cognitive performance in psychosis: A cross-sectional study in patients with non-affective psychotic illness and their unaffected siblings. *Psychological Medicine, 42*(4), 705–716. doi: 10.1017/s0033291711001656

Negriff, S., Fung, M. T., & Trickett, P. K. (2008). Self-rated pubertal development, depressive symptoms and delinquency: Measurement issues and moderation by gender and maltreatment. *Journal of Youth and Adolescence, 37,* 736–746.

(Continued)

(Continued)

Oh, S., Salas-Wright, C. P., & Vaughn, M. G. (2018). Trends in depression among low-income mothers in the United States, 2005–2015. *Journal of Affective Disorders, 235*, 72–75. doi: https://doi.org/10.1016/j.jad.2018.04.028

Schwarzer, R., & Hallum, S. (2008). Perceived teacher self-efficacy as a predictor of job stress and burnout. *Applied Psychology: An International Review, 57* (Suppl. 1), 152–171.

Tarantino, N., Brown, L. K., Whiteley, L., Fernández, M. I., Nichols, S. L., & Harper, G. (2018). Correlates of missed clinic visits among youth living with HIV. *AIDS Care, 30*(8), 982–989. doi: 10.1080/09540121.2018.1437252

Consensus Statements

Consensus statements are common in health and medicine and provide guidance to physicians and patients on how to identify and care for dozens of problems, including knee replacement, epilepsy, and cataracts. A group or panel of knowledgeable individuals issues consensus statements, and they do so because the available literature on that topic is incomplete or contradictory. Consensus panels generally agree that the best way to get information is through controlled experimentation. But good study data are not always immediately available.

The best consensus statements result from a consideration of the world's literature in combination with group process methods known to make maximum use of participants' expertise. The number of participants in most consensus development activities ranges from 9 to 14. The most famous are the National Institutes of Health (NIH) Consensus Statements (https://consensus.nih.gov).

Reviewers of the literature are often tempted to include consensus statements to justify the need for a study or its conclusions. Consensus statements are observational studies, however, and are prey to many of the same limitations as observational studies are.

Books

Many books contain excellent literature reviews, and their bibliographies are a gold mine for other reviewers. Books are also essential guides to understanding theory and for helping you to validate the need for your study, confirm your choice of literature, and certify (or contradict) its findings. By definition, however, literature reviews are based on an analysis of the original studies. Original studies allow the reviewer to report, "Jones and Smith *found* . . ." With books, you must report, "Jones and Smith *say* . . ."

Internal and External Validity

A study design with **external validity** produces results that apply to the study's target population. An externally **valid** survey of the preferences of airline passengers over 45 years of age means that the findings apply to all airline passengers of that age.

A design is internally valid if it is free from nonrandom error or bias. A study design must be internally valid in order to be externally valid. One of the most important questions to ask when reviewing the literature is, Does this study's design have **internal validity**? The following is a checklist of the influences on a study that threaten its internal validity.

Internal Invalidity: A Checklist of Potential Threats to a Study's Accuracy

✓ **Maturation** **Maturation** refers to changes within individuals that result from natural, biological, or psychological development. For example, in a 5-year study of a preventive health education program for high school students, the students may mature intellectually and emotionally, and this new maturity may be more important than the program in producing changes in health behavior.

✓ **Selection** *Selection* refers to how people were chosen for a study and, if they participate in an experiment, how they were assigned to groups. Selection bias is minimized when every eligible person or unit has an equal, nonzero chance of being included.

✓ **History** Historical events may occur that can bias the study's results. For example, suppose a national campaign has been created to encourage people to make use of preventive health care services. If a change in health insurance laws favoring reimbursement for preventive health care occurs at the same time as the campaign, it may be difficult to separate the effects of the campaign from the effects of increased access to care that have been created by more favorable reimbursement for health care providers.

✓ **Instrumentation** Unless the **measures** used to collect data are dependable, you cannot be sure that the findings are accurate. For example, in a **pretest, posttest, or self-controlled design**, an easier measure after an intervention or program than before one will erroneously favor the intervention.

✓ **Statistical regression** Suppose people are chosen for an intervention to foster tolerance. The basis for selection, say, was their extreme views, as measured by a survey. A second administration of the survey (without any intervention) may appear to suggest that the views were somehow softened, but in fact, the results may be a statistical artifact. This is called regression to the mean.

✓ **Attrition (loss to follow-up)** *Attrition* is another word for loss of data such as occurs when participants do not complete all or parts of a survey. People may not complete study activities because they move away, become ill or bored, and so on. Sometimes participants who continue to provide complete data throughout a long study are different from those who do not, and this difference biases the findings.

Risks to external validity are most often the consequence of the way in which participants or **respondents** are selected and assigned. For example, respondents in an experimental situation may answer survey questions atypically because they know they are in a special experiment; this is called the **"Hawthorne" effect**. External validity is also a risk just because respondents are tested, surveyed, or observed. They may become alert to the kinds of behaviors that are expected or favored. Sources of external invalidity are included in the following checklist.

External Invalidity: A Checklist of Risks to Avoid

✓ **Reactive effects of testing** A measure given before an intervention can sensitize participants to its aims. Suppose two groups of junior high school students are eligible to participate in a program to teach **ethics**. Say that the first group is surveyed regarding its perspectives on selected ethics issues and then shown a film about young people from different backgrounds faced with ethical dilemmas. Suppose that the second group of students is just shown the film. It would not be surprising if the first group performed better on a postmeasure if only because the group was sensitized to the purpose of the movie by the questions on the "premeasure."

✓ **Interactive effects of selection** This threat occurs when an intervention or program and the participants are a unique mixture—one that may not be found elsewhere. Suppose a school volunteers to participate in an experimental program to improve the quality of students' leisure time activities. The characteristics of the school (some of which may be related to the fact that it volunteered for the experiment) may interact with the program so that the two together are unique; the particular blend of school and intervention can limit the applicability of the findings.

✓ **Reactive effects of innovation** Sometimes the environment of an experiment is so artificial that all who participate are aware that something special is going on and behave uncharacteristically.

✓ **Multiple-program interference** It is sometimes difficult to isolate the effects of an experimental intervention because of the possibility that participants are in other complementary activities or programs.

The following examples illustrate how internal and external validity are affected in two different study designs.

How the Choice of Research Design Affects Internal and External Validity

1. Parallel Controls Without Random Assignment

   Description. The Work and Stress Program is a yearlong program to help reduce on-the-job stress. Eligible people can enroll in one of two variations of the program. To find out if participants are satisfied with the quality of the program, both groups complete an in-depth questionnaire at the end of the year, and the results are compared.

   Comment. The internal validity is potentially marred by the fact that the participants in the groups may be different from one another at the beginning of the program. More "stressed" persons may choose one program over the other, for example. Also, because of initial differences, the attrition, or loss to follow-up, rate may be affected. The failure to create randomly constituted groups will jeopardize the study's external validity by the interactive effects of selection.

(Continued)

(Continued)

2. Parallel Controls With Randomization

   Description. Children's Defense Trust commissioned an evaluation of three different interventions to improve school performance. Eligible children were randomly assigned to one of the three interventions, baseline data were collected, and a 3-year investigation was made of effectiveness and efficiency. At the end of the 3 years, the children were examined to determine their functioning on a number of variables, including school performance and behavior at home and at school. The children were also interviewed extensively throughout the study.

   Comment. This design is internally valid. Because children were randomly assigned to each intervention, any sources of change that might compete with the intervention's impact will affect all three groups equally. To improve external validity, the findings from a study of other children will be compared with those from the Children's Defense Trust. This additional comparison does not guarantee that the results will hold for a third group of children. Another consideration is that school administrators and staff may not spend as much money as usual because they know the study involves studying efficiency (reactive effects of innovation). Finally, we do not know if and how baseline data collection affected children's performance and interviews (interaction between **testing** and the intervention).

Criterion for Quality: Sampling

What Is a Sample?

A *sample* is a portion or subset of a larger group called a population.

The target population consists of the institutions, persons, problems, and systems to which or to whom a study's findings are to be applied or *generalized.* Consider these two target populations and samples.

Two Target Populations and Two Samples

1. Target population: All teacher training programs in the state

   Program. Continuous Quality Improvement: An intervention to monitor and change the quality of teacher training. One index of quality is the performance of students on statewide reading and math tests.

   Sample. Five teacher training institutions were selected to try out the Quality Improvement experiment. After 1 year, for all participating teacher trainees, a 10% sample of student performance in reading and math was evaluated.

   Comment. The target for this study is all teacher training programs in the state. Five will be selected for a Continuous Quality Improvement program. To appraise the program's quality, the researcher sampled 10% of students to assess their performance in reading and math. The findings were applied to all teacher training programs in the state.

2. Target population: All students needing remediation in reading

   Program. Options for Learning

   Sample. Five schools in three counties; within each school, 15 classes; for each class, at least two to five students who need remediation in reading.

   Comment. Students who need assistance in reading were the targets of the program. The researchers selected five schools in three counties and, within them, 15 classes with two to five students in each. The findings were applied to all students who need special aid in reading.

Inclusion and Exclusion Criteria or Eligibility of Participants

A sample is a constituent of a larger population to which a study's findings will be applied. If a study plans to investigate the impact of a counseling program on children's attitudes toward school, for example, and not all students in need of more favorable attitudes are to be included in the program, then the researcher has to decide on the types of students who will be the focus of the study. Will the research concentrate on students of a specific age? With particular achievement levels? With poor attendance records?

From the literature reviewer's perspective, one mark of methodological quality is evidence of explicit inclusion and exclusion criteria. Failure to be explicit means that the reviewer will find it practically impossible to determine who was included and excluded from the study and for whom the findings are appropriate. Claims made by researchers regarding the applicability of their study's findings to groups of people or places can be evaluated only within the context of the subjects or participants who were eligible to be in a study and who actually participated.

The next example contains hypothetical inclusion and exclusion criteria for an evaluation of such a program to foster children's favorable attitudes toward school.

Inclusion and Exclusion Criteria for a Study of the Impact of a Program to Foster Favorable Student Attitudes to School

Inclusion

- All students attending schools in the zip codes listed below (not included in this example) who are currently in the sixth through ninth grade
- Students who speak English or Spanish
- Students who have participated in the E.T. (Eliminate Truancy) program

Exclusion

- All students who are currently incarcerated

Comment. The researcher set explicit criteria for the sample of students who are included in the study and for whom its findings are appropriate. The sample includes children in the sixth through ninth grade who speak English and Spanish, live within the confines of certain zip codes, and have participated in the Eliminate Truancy (ET) program. The findings are not applicable to students who meet just some of the criteria; for example, they are in the sixth grade, live in one of the specified zip codes, speak Spanish, but have NOT participated in the ET program.

Methods of Sampling

Sampling methods are usually divided into two types. The first is called *random* or *probability sampling*, and it is considered the best way to ensure the validity of any inferences made about a program's effectiveness and its **generalizability**. In probability sampling, every member of the target population has a known probability of being included in the sample. Probability or random sampling methods sometimes require knowledge of probability statistics; many statistical software programs have random-sampling capabilities, but their use is not meant for the statistically challenged.

A second type of sampling method produces a *convenience sample*. A convenience sample consists of participants who are selected because they are available. In convenience sampling, some members of the target population have a chance of being chosen, but others do not because they are not present when the sample is assembled. As a result, the data collected from a convenience sample may not be applicable to the target group at all. (The people who show up may differ from those who do not.) For example, suppose a researcher who is concerned with evaluating a college's student health service decided to interview 100 students who came for assistance during the week of December 26 to January 1. Suppose that 100 students are interviewed. The problem is that the end of December in some parts of the world is associated with respiratory viruses and skiing accidents; moreover, many schools are closed during that week, and students are not around. Thus, the resulting data could very well be biased because the survey excluded many students simply because they were not on campus (and if they were ill, did not receive care or received care elsewhere).

Simple Random Sampling

In simple random sampling, every subject or unit has an equal chance of being selected. Because of this equality of opportunity, random samples are considered relatively unbiased. Typical ways of selecting a simple random sample include using a table of random numbers or a computer-generated list of random numbers and applying them to lists of prospective participants.

Suppose a researcher wanted to use a table and had the names of 20 psychologists from which 10 were to be selected at random. The list of names is called the sampling frame. First, the researcher would assign a number to each name, 1 to 20 (e.g., Adams = 1; Baker = 2; Thomas = 20). Then using a table of random numbers, which can be found online (enter table of random numbers or digits) and in many statistics books, the researcher would choose the first 10 digits between 1 and 20. Or a list of 10 numbers between 1 and 20 can be generated using any one of the most commonly available statistical programs.

Systematic Sampling

Suppose a researcher had a list with the names of 3,000 high school seniors from which a sample of 500 was to be selected. In systematic sampling, 3,000 would be divided by 500 to yield 6, and every sixth name would be selected. An alternative would be to select a number at random, say, by tossing dice. Suppose a toss came up with the number 5. Then, the fifth name would be selected first, then the 10th, 15th, and so on until 500 names were selected.

Systematic sampling should not be used if repetition is a natural component of the sampling frame or list from which the sample is to be drawn. For example, if the frame is a list of names, those beginning with certain letters of the alphabet might get excluded because, for certain ethnicities, they appear infrequently.

Stratified Sampling

A stratified random sample is one in which the population is divided into subgroups or "strata," and a random sample is then selected from each group. For example, in a program to teach students problem-solving skills, a researcher might choose to sample students of differing age, achievement, and self-confidence. Age, achievement, and self-confidence are the strata.

The strata or subgroups are chosen because the researcher provides evidence that they are related to the dependent variable or outcome measure—in this case, problem-solving skills. That is, the researcher provides the reviewer with convincing data—from high-quality literature and expert opinion—that age, general achievement, and perceptions of self-confidence influence ability to problem solve.

If the researcher neglects to use stratification in the choice of a sample, the results may be confounded. Suppose the literature suggests that women of varying ages react differently to a certain type of health initiative. If the researcher fails to stratify by age, good and poor performance may be averaged among the women participating in the initiative, and no effect will be seen—even if one or more groups benefited.

When stratification is not used, statistical techniques (such as **analysis of covariance** and regression) may be applied retrospectively (after the data have already been collected) to correct for confounders ("**covariates**") of the dependent variables or outcomes. In general, it is better to anticipate confounding variables by sampling prospectively than to correct for them by analysis, retrospectively. The reason is that statistical corrections require very strict assumptions about the nature of the data, assumptions for which the sampling plan may not have been designed. With few exceptions, using statistical corrections afterward results in a loss of **power** or ability to detect true differences.

Cluster Sampling

Clusters are naturally occurring groups such as schools, clinics, community-based service organizations, cities, states, and so on. In cluster sampling, the population is divided into batches. The batches can be randomly selected and assigned, and their constituents can be randomly selected and assigned. For example, suppose that 10 counties are trying out a new program to improve voter registration; the control program is the traditional program. With random cluster sampling, each county is a cluster, and each can be selected and assigned at random to the new or traditional program.

Convenience Sampling

Convenience samples are those for which the probability of selection is unknown. Researchers use convenience samples because they are easy to get. This means that some people have no chance at all of being selected, simply because they are not around to be chosen. These samples are considered biased, or not representative of the target population, unless proven otherwise (through statistical methods, for example).

The Sampling Unit

A major concern in sampling is the potential discrepancy between the "unit" to be sampled and the unit that is analyzed statistically. For instance, suppose a group of researchers is interested in finding out about patient satisfaction in a medical organization that has five large clinics. The researchers survey 6,000 patients in a clinic in the far north and 5,000 in a clinic in the far south. On the basis of the results in both clinics, the researchers report that patients in the medical organization are extremely satisfied with their medical care. The findings show, for instance, that of the 11,000, nearly 98% state that their care is as good as or better than any care they have ever received. The medical care organization is very pleased with these findings.

The literature reviewer has to be careful with the conclusion of studies that do not address discrepancies between who is sampled and whose data are analyzed. In the above example, two clinics were sampled (the sampling unit), but data were analyzed for 11,000 patients (the analysis unit). Because only two of five clinics were in the sample, you cannot be sure that the two clinics are not different from the remaining three and that you have a sample size of 2 and not of 11,000. A better strategy, but one that is much more difficult to implement, might have been to sample 11,000 persons across all five clinics.

Statistical methods are available for "correcting" for the discrepancy between units of sampling and analysis. When appropriate, examine if and how discrepancies between sampling and analysis units are handled. Because the analysis methods used to correct for clustering are complex, you may need statistical consultation.

The Size of the Sample

The size of the sample is important for several reasons. Small samples may not be able to include the mix of people or programs that should be included in a study and may be unable to detect an **effect** even if one would have taken place with more people. A study's ability to detect an effect is its power. A **power analysis** is a statistical method of identifying a sample size that is large enough to detect the effect, if one actually exists. A most commonly used research design is one in which two randomly assigned groups are compared to find out if differences exist between them. "Does Program A differ from Program B in its ability to improve satisfaction? Quality of life? Reading? Math? Art? Mental health? Social functioning?" is a fairly standard research question. To answer the question accurately, the researcher has to design the study so that a sufficient number of subjects are in each program group so that if a difference is actually present, it will be uncovered. Conversely, if there is no difference between the two groups, the researcher does not want to conclude falsely that there is one.

Statistical methods are available for researchers to identify a sample that is large enough to detect actual effects. The power of an experimental study is its ability to detect a true difference—in other words, to detect a difference of a given size (say, 10%) if the difference actually exists. Many published articles do not include their power calculations, so if differences are not observed, the problem may have been that the sample was not large enough to detect a difference among groups, even if one may have been present.

Response Rate

The **response rate** is the number who are measured, observed, or responded (numerator) divided by the number of eligible respondents (denominator):

$$\text{Response Rate} = \frac{\text{Number who respond}}{\text{Eligible to respond}}$$

All studies aim for a high response rate. No standard exists, however, to assist the literature reviewer in deciding whether the aim was achieved and, if not, the effect on the study's outcomes.

Consider two examples. In the first, 50% of eligible persons complete all **items** on a health survey. In the second, 100% of eligible persons respond, but they fail to complete about 50% of the items on the survey.

Nonresponse: Subjects and Items

1. The National State Health Interview is completed by 50% of all who are eligible. Health officials conclude that the 50% who do not participate probably differ from participants in their health needs and demographic characteristics.
2. According to statistical calculations, the Commission on Refugee Affairs (CORA) needs a sample of 100 for their mailed survey. Based on the results of previous mailings, a refusal rate of 20% to 25% is anticipated. Just in case, 125 eligible people are sent a survey. One hundred twenty persons respond, but on average, they answer fewer than half of all questions.

In the first case described above, 50% of eligible state residents do not complete the interview. These nonrespondents may be very different in their health needs, incomes, and education than the 50% who do respond. When nonrespondents and respondents differ on important factors, this is called nonresponse bias. Nonresponse bias may seriously impair a study's generalizability (external validity) because the findings, which were expected to apply to a relatively broad group, are now applicable just to the persons who responded or agreed to participate. Reviewers should be on the alert for studies that do not explain the consequences of nonresponse. Questions such as these should be answered: Of those who were eligible, how many participated? What was the reason for the nonresponse? How do responders compare to nonresponders? How is the study's internal and external validity affected by the nonresponse?

In addition to person nonresponse, item nonresponse may introduce bias. Item nonresponse occurs when respondents do not complete all items on a survey or test. This type of bias comes about when respondents do not know the answers to certain questions or refuse to answer them because they cannot (e.g., they do not understand the questions) or believe them to be sensitive, embarrassing, or irrelevant.

Statistical methods may be used to correct for nonresponse to the entire survey or to just some items. One method involves "weighting." Suppose a survey wants to compare younger (younger than 25 years) and older (26 years and older) college students' career goals. A review of school records reveals that younger students are 40% of the population. Although all 40% are given a survey to complete, only 20% do so. Using statistical methods, the 20% response rate can be weighted to become the equivalent of 40%. The accuracy of the result depends on the younger respondents being similar in their answers to the nonrespondents and different in their answers to the older respondents.

Another method of correcting for nonresponse is called imputation. With imputation, values are assigned for the missing response, using the responses to other items as supplementary information. Scientifically sound studies explain in detail how missing data are handled and the effects of missing data on the findings.

The following checklist can be used when reviewing a study's presentation and quality as it pertains to research design and sampling. The list is probably too extensive to use for any single literature review, and so you must decide which questions to answer on a case-by-case basis.

A Checklist for Evaluating the Presentation and Quality of Study Design and Sampling

✓ If more than one group is included in the study, are the participants randomly assigned to each?

✓ Are participants measured over time? If so, is the number of observations explained? Justified?

✓ If observations or measures are made over time, are the choice and effects of the time period explained?

✓ Are any of the participants "blinded" to the group—experimental or control—to which they belong?

- ✓ If historical controls are used, is their selection explained? Justified?
- ✓ Are the effects on internal validity of choice, equivalence, and participation of the sample subjects explained?
- ✓ Are the effects on external validity (generalizability) of choice, equivalence, and participation of the subjects explained?
- ✓ If a sample is used, are the subjects randomly selected?
- ✓ If the unit sampled (e.g., students) is not the population of main concern (e.g., teachers are), is this addressed in the analysis or discussion?
- ✓ If a sample is used with a nonrandom sampling method, is evidence given regarding whether they are similar to the target population (from which they were chosen) or to other groups in the study?
- ✓ If groups are not equivalent at baseline, is this problem addressed in analysis or interpretation?
- ✓ Are criteria given for including subjects?
- ✓ Are criteria given for excluding subjects?
- ✓ Is the sample size justified (say, with a power calculation)?
- ✓ Is information given on the size and characteristics of the target population?
- ✓ If stratified sampling is used, is the choice of strata justified?
- ✓ Is information given on the number and characteristics of subjects in the target population who are eligible to participate in the study?
- ✓ Is information given on the number and characteristics of subjects who are eligible and who also agree to participate?
- ✓ Is information given on the number and characteristics of subjects who are eligible but refuse to participate?
- ✓ Is information given on the number and characteristics of subjects who dropped out or were lost to follow-up before completing all elements of data collection?
- ✓ Is information given on the number and characteristics of subjects who completed all elements of data collection?

- ✓ Is information given on the number and characteristics of subjects on whom some data are missing?
- ✓ Are reasons given for missing data?
- ✓ Are reasons given why individuals or groups dropped out?

SUMMARY OF KEY POINTS

An efficient literature search is always filtered through two screens. The first screen is primarily practical. It is used to identify studies that are potentially pertinent in that they cover the topic, are in a language you read, and appear in a publication you respect. The second screen is for methodological quality, which is used to identify the best available studies in terms of their adherence to the methods that scientists and scholars rely on to gather sound evidence. You must use both screens to ensure the review's relevance and accuracy.

- Typical practical criteria for literature review searches include the following:

  Publication language

  Journal

  Author

  Setting

  Participants

  Type of program or intervention

  Research design

  Sampling

  Date of publication

  Date of data collection

  Duration of data collection

  Content (topics, variables)

  Source of financial support

- *Methodological quality* refers to how well—scientifically—a study has been designed and implemented to achieve its objectives. The highest quality studies adhere to rigorous research standards.
- A study's research design refers to the way in which its subjects or constituents—students, patients, and customers—are organized and observed. Research designs are traditionally categorized as experimental or observational.
- Typical experimental designs include the following:

  Parallel controls in which groups are assigned randomly or the true experiment. *Parallel* means that each group is assembled at the same time. When 500 students are randomly assigned to an experimental group while, at the same time, 500 are assigned to a control group, you have parallel controls (each group is assembled at the same time) with random assignment. This design is also called a simple **randomized controlled trial** or true experiment.

  Parallel controls in which participants are not randomly assigned to groups or the quasi-experiment. These are called nonrandomized controlled trials, quasi-experiments, or nonequivalent controls. When children are assigned to an experimental after-school program because they live in City A, and another group is assigned to a control program because they live in City B, you have a quasi-experiment or nonrandomized trial.

  Self-control. These require premeasures (also called pretests) and postmeasures (also called **posttests**) and are also called longitudinal or before-after or pretest-posttest designs. For instance, a study is longitudinal if employees in a fitness program are given a series of physical examinations before participation in a new health promotion program and 6 months, 1 year, and 2 years after participation.

  Historical controls. These use "**normative data**" against which to compare a group. Normative data are historical because they come from already existing databases. For instance, a researcher who evaluates a program to improve employees' blood pressure levels and uses standard tables of normal blood pressure to monitor improvement is conducting a study that uses historical controls.

- Observational designs produce information on groups and phenomena that already exist. Researchers who do observational studies have "less control" than do researchers who conduct experimental studies. Because of this, observational designs are considered less rigorous than experimental research designs. Typical observational designs include the following:

Cohorts. These designs provide data about changes in a specific population. Suppose a survey of the aspirations of athletes participating in the Olympics is given in 2000, 2004, and 2008. This is a cohort design, and the cohort is 2000 Olympians.

Case controls. These studies help explain a current phenomenon. At least two groups are included. When you survey the medical records of a sample of people with heart disease and a sample without the disease to find out about the similarities and differences in past illnesses, you have used a case control design.

Cross-sections. These provide descriptive or survey data at one fixed point in time. A survey of American voters' current choices is an example of a cross-sectional research design.

- A study design is internally valid if it is free from nonrandom error or bias. A study design must be internally valid to be externally valid and to produce accurate findings. One of the most important questions to ask when reviewing the literature is this: Does this study's design have internal validity? Threats to internal validity include the following:
  - ✓ **Maturation**. *Maturation* refers to changes within individuals that result from natural, biological, or psychological development.
  - ✓ **Selection**. *Selection* refers to how people were chosen for the study and, if they participated in an experiment, how they were assigned to groups.
  - ✓ **History**. Historical events are extraneous forces that occur while the study is in operation and may interfere with its implementation and outcomes.
  - ✓ **Instrumentation**. Unless the measures used to collect data are dependable or reliable, the findings are unlikely to be accurate.
  - ✓ **Statistical regression**. A tendency of very high or low values to move toward the mean or average: A statistical artifact.
  - ✓ **Attrition**. This is another word for loss of data such as occurs when participants do not complete all or parts of the study's data collection instruments.
- A study design with external validity produces results that apply to the study's target population.
- Threats to external validity are most often the consequence of the way in which participants or respondents are selected and assigned. For example,

respondents in an experimental situation may answer questions atypically because they know they are in a special experiment. External validity is also at risk just because respondents are tested, surveyed, or observed. They may become alert to the kinds of behaviors that are expected or favored. Threats to external validity include the following:

- ✓ **Reactive effects of testing**. A premeasure can sensitize participants to the aims of an intervention.
- ✓ **Interactive effects of selection**. This occurs when an intervention and the participants are a unique mixture—one that may not be found elsewhere.
- ✓ **Reactive effects of innovation**. Sometimes the environment of an experiment is so artificial that all who participate are aware that something special is going on and behave uncharacteristically.
- ✓ **Multiple-program interference**. It is sometimes difficult to isolate the effects of an experimental intervention because of the possibility that participants are in other complementary activities or programs.

- Sampling methods are usually divided into two types. The first is called probability sampling, and it is considered the best way to ensure the validity of any inferences made about a program's effectiveness and its generalizability. In probability sampling, every member of the target population has a known probability of being included in the sample. Few studies use true probability sampling. The second type of sample is the convenience sample in which participants are selected because they are available. In convenience sampling, some members of the target population have a chance of being chosen, but others do not. As a result, the data that are collected from a convenience sample may not be applicable to the target group at all.
- Types of sampling include the following:

Simple random sampling. In simple random sampling, every subject or unit has an equal chance of being selected. Because of this equality of opportunity, random samples are considered relatively unbiased.

Systematic sampling. Suppose a researcher had a list with the names of 3,000 high school seniors from which a sample of 500 was to be selected. In systematic sampling, 3,000 would be divided by 500 to yield 6, and every sixth name would be selected.

Stratified sampling. A stratified random sample is one in which the population is divided into subgroups or "strata," and a random sample is then selected from each group.

Cluster sampling. A cluster is a naturally occurring group such as schools, clinics, community-based service organizations, cities, states, and so on. In cluster sampling, the population is divided into batches. The batches can be randomly selected and assigned, and their constituents can be randomly selected and assigned.

Convenience samples. Convenience samples are those for which the probability of selection is unknown. Researchers use convenience samples because they are easy to get. This means that some people have no chance at all of being selected, simply because they are not around to be chosen. These samples are considered biased or not representative of the target population, unless proven otherwise (through statistical methods, for example).

- A study's ability to detect an effect if it is present is its power. A power analysis is a statistical method of identifying a sample size that is large enough to detect the effect, if one actually exists.
- The response rate is the number who respond (numerator) divided by the number of eligible respondents (denominator):

$$\text{Response Rate} = \frac{\text{Number who respond}}{\text{Number Eligible to respond}}$$

EXERCISES

1. The Community Family Center had 40 separate counseling groups, each with about 30 participants. The director of the center conducts and reports on an experiment to improve attendance rates at the sessions. Random selection of individuals from all group members for the experiment was impossible; such selection would have created friction and disturbed the integrity of some of the groups. Instead, a design was used in which five of the groups—150 people—were randomly selected to take part in the experiment, and five continued to receive traditional counseling. Every 3 months, the director compares the attendance of all persons in the experimental group with those in the control group.

Compare and comment on the sampling and analysis units.

a. Which method of sampling is used?

b. Compare and comment on the units of sampling and analysis.

2. The Medical Group developed an interactive computer-based educational intervention to prevent strokes. A study was conducted to compare the computer intervention with the traditional method that consisted of written handouts routinely given to all persons between 45 and 75 years of age. The study was experimental, with parallel controls. Of 310 eligible persons, 140 were between 45 and 60 years old, and 62 of these were men. The remaining 170 were between 61 and 75 years, and 80 of these were men. The researchers randomly selected 40 persons from each of the four subgroups and randomly assigned every other person to the computer intervention and the remainder to the control (written materials).

   a. Which sampling method is used?

   b. Which eligibility criteria do you think may have applied?

   c. Draw the sampling plan.

3. Two hundred teen counselors signed up for a continuing education program. Only 50, however, participated in an evaluation of the program's impact. Each participant was assigned a number from 001 to 200 and, using a table, 50 names were selected by moving down columns of three-digit random numbers and taking the first 50 numbers within the range 001 to 200.

   a. Which sampling method is used?

   b. What is the response rate?

4. What is the research design in the following studies? What are the threats to internal and external validity?

   Study A. The ABC MicroLink Company experimented with a program to provide its employees with options for caring for their older parents. Human resources staff interviewed all employees to find out how much they learned from the program and if they were planning to use its content.

   Study B. Teens in the ALERT program voluntarily joined one of three 1-month rehabilitation programs. Data were collected on teens' knowledge and self-confidence before and after participation in each program.

ANSWERS

1a. Cluster sampling

1b. The sampling unit was a "group," so there were five units or groups. The analysis compared average attendance among 150 persons in the experiment and 150 in the control. A problem with the accuracy of the results may arise if one or more of the groups has a unique identity (e.g., more cohesive, more cooperative, more knowledgeable).

2a. Stratified random sampling

2b. Must be between 45 and 75 years of age. Must be willing to use interactive computer for educational purposes.

2c. Sampling plan

The Population

	Age	
	45–60	61–75
Men	62	80
Women	78	90

The Sample

	Age	
	45–60	61–75
Men	40	40
Women	40	40

3a. Simple random sampling

3b. 50/200 or 25%

4. *Study A.* Cross-sectional design. *The internal validity* of cross-sectional designs may be affected by nearly all possible threats. Historical events (such as new legislation regarding the care of older parents), for example, may occur at the same time as the program, and these may be as or more influential than the program. Selection may threaten internal validity because of the nature of the sample that participates and completes all study activities. Because of the perilous state of internal validity in cross-sectional designs, you cannot count on them to produce externally valid results.

External validity may be influenced by the reactive effects of innovation.

Study B. Cohort design. Selection is a possible risk to *internal validity* because participants in the two groups may have been different from one another at the beginning of the program. For example, more self-confident teens may choose one program over the other. Also, attrition may be different between the two groups. Risks to *external validity* include the reactive effects of innovation, interactive effects of selection, and, possibly, multiple-program interference.

SUGGESTED READINGS

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental design for research.* Chicago, IL: Rand-McNally.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Thousand Oaks, CA: Sage.

De Vaus, D. (2002). *Research design in social research.* Thousand Oaks, CA: Sage.

Fink, A. (2008). *Practicing research: Discovering evidence that matters*. Thousand Oaks, CA: Sage

Fink, A. (2012). *Evidence-based public health*. Thousand Oaks, CA: Sage

Henry, G. T. (1990). *Practical sampling.* Newbury Park, CA: Sage.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D., Hearst, N., & Newman, T. B. (Eds.). (2006). *Designing clinical research* (2nd ed., Chaps. 5 and 6). Philadelphia, PA: Lippincott, Williams, & Wilkins.

McIntyre, A. (2008). *Participatory action research.* Thousand Oaks, CA: Sage.

Riegelman, R. K., & Hirsch, R. P. (2004). *Studying a study and testing a test: How to read the health science literature.* Boston, MA: Little, Brown.

For sample size software, go to your favorite search engine and type in *sample size calculator.*

For a tool to measure risk of bias in randomized controlled trials, use the Cochrane Collaboration's risk assessment tool:

https://www.ncbi.nlm.nih.gov/books/NBK132494/bin/appf-fm1.pdf