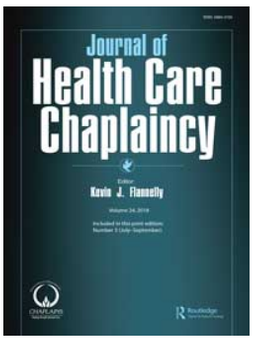

Journal of Health Care Chaplaincy

ISSN: 0885-4726 (Print) 1528-6916 (Online) Journal homepage: https://www.tandfonline.com/loi/whcc20

Threats to the Internal Validity of Experimental and Quasi-Experimental Research in Healthcare

Kevin J. Flannelly, Laura T. Flannelly & Katherine R. B. Jankowski

To cite this article: Kevin J. Flannelly, Laura T. Flannelly & Katherine R. B. Jankowski (2018) Threats to the Internal Validity of Experimental and Quasi-Experimental Research in Healthcare, Journal of Health Care Chaplaincy, 24:3, 107-130, DOI: 10.1080/08854726.2017.1421019

To link to this article: https://doi.org/10.1080/08854726.2017.1421019

Journal of Health Care Chaplaincy, 24:107–130, 2018

ISSN: 0885-4726 print/1528-6916 online
DOI: 10.1080/08854726.2017.1421019

Threats to the Internal Validity of Experimental and Quasi-Experimental Research in Healthcare

KEVIN J. FLANNELLY and LAURA T. FLANNELLY
Center for Psychosocial Research, Massapequa, New York, USA

KATHERINE R. B. JANKOWSKI
Center for Psychosocial Research, Massapequa, New York, USA; Iona College, New Rochelle, New York, USA

The article defines, describes, and discusses the seven threats to the internal validity of experiments discussed by Donald T. Campbell in his classic 1957 article: history, maturation, testing, instrument decay, statistical regression, selection, and mortality. These concepts are said to be threats to the internal validity of experiments because they pose alternate explanations for the apparent causal relationship between the independent variable and dependent variable of an experiment if they are not adequately controlled. A series of simple diagrams illustrate three pre-experimental designs and three true experimental designs discussed by Campbell in 1957 and several quasi-experimental designs described in his book written with Julian C. Stanley in 1966. The current article explains why each design controls for or fails to control for these seven threats to internal validity.

KEYWORDS experimentation, internal validity, methodology, quasi-experiments, research design

Address correspondence to Kevin J. Flannelly, PhD, Center for Psychosocial Research, 33 Maple Street, Massapequa, NY 11758, USA. E-mail: kjflannelly@gmail.com

INTRODUCTION

Claude Bernard, the father of experimental medicine, proposed that experimentation was needed in medicine to help physicians "conserve health and cure disease" (Bernard, 1865/1957, p. 1), by providing knowledge about the causes of normal states of health (i.e., anatomy and physiology), the causes of disease (i.e., pathology), and the effectiveness of therapeutic treatments. Bernard's 1865 book, *An Introduction to the Study of Experimental Medicine,* emphasized the difference between the simple observation of natural changes in physiological processes over the course of time and experimentation, in which the researcher intervenes in some way to change the natural course of physiological processes. There are four important stages in the research process according to Bernard. The researcher (a) observes a natural phenomenon, (b) develops a hypothesis about the phenomenon, (c) applies a procedure to test this hypothesis, and (d) compares the results before and after applying the procedure.

Six years earlier, the philosopher John Stuart Mill (1859) published a book that laid out the principles he mainly developed for inferring causality from research results. The first principle was that the putative or presumed cause must occur before the effect. The second was that the effect must always occur when the presumed cause occurs. Third, the effect must not occur when the presumed cause is absent. Fourth, the presumed cause must be isolated from other potential causes of the effect. Fifth, the presumed cause must be produced artificially to ensure that the cause is isolated from all other potential causes. These principles set the standard for establishing casual relationships between presumed causes and effects (see K. J. Flannelly & Jankowski, 2014a). The fourth and fifth principles distinguish experimentation from other forms of research in that the experimenter must be able to manipulate the presumed cause and isolate the effects of the presumed cause from other events or variables (i.e., see discussion of variables by L. T. Flannelly, Flannelly, & Jankowski, 2014a) that could cause the observed effect.

Based on Mill's (1859) principles, an experimenter tries to control or hold constant all the variables that can affect the outcome (the dependent variable) of an experiment apart from the experimental manipulation (Keppel & Wickens, 2004), which is also called the independent variable, intervention, or treatment (see L. T. Flannelly et al., 2014a). The variables that the researcher wants to control are known as both extraneous variables, because they are extraneous to the purpose of the experiment, and confounding variables, because their effects are confounded with the effects of the independent variable if they are not properly controlled. The explicit concern is that "the operation of some extraneous variable causes the observed values of the dependent variable to inaccurately reflect the effect of the independent variable" (Cherulnik, 1983, p. 21). In other words, the

observed effect of the experiment is not due to the independent variable, but to the extraneous variable. Thus, the failure to control extraneous variables undermines the ability of researchers to logically make the causal inference that the apparent effect of an experimental manipulation is, in fact, the result of the manipulation (i.e., the independent variable or intervention). Unfortunately, it is not very easy to control extraneous variables outside of a laboratory setting (Rubinson & Neutens, 1987).

Donald T. Campbell published a classic article in 1957 that describes the types or classes of extraneous variables a researcher must control in order to be able to make causal inferences from an experiment. He also explains why causal inferences cannot be made from various types of experimental designs. Campbell called the degree to which a design controls extraneous variables and, thus, permits causal inferences to be made regardiing the association between the independent and dependent variable, the experiment's "internal validity." An experiment with a high degree of internal validity has reduced the potential influence of extraneous variables to such an extent that the independent variable is the most likely cause of the observed change in the dependent variable. An experiment with low internal validity has not eliminated the possibility that some variable other than the independent variable has caused the observed change in the dependent variable. Even though Campbell's article is titled "Factors Relevant to the Validity of Experiments in Social Settings," it is applicable to all experimentation involving human participants.

Campbell's (1957) article focuses on seven classes of extraneous variables that can undermine the internal validity of an experiment if they are not controlled by the experimental design of a study. Thus, these classes of extraneous variables are called "threats to internal validity." Campbell named them: *history, maturation, testing, instrument decay, statistical regression, selection,* and *mortality*. Properly controlling for these variables eliminates them as rival explanations for the results of an experiment. Campbell also discussed factors that affect the "external validity" or generalizability of an experiment's results, but we will not address them in this article.

Campbell's (1957) analysis of the threats to internal validity posed by different designs was extended in his later writings that expanded from pre-experimental and experimental designs to include what he called quasi-experimental designs (Campbell & Stanley, 1966; Cook & Campbell, 1979), which are discussed later. Although the concepts of internal validity and threats to the internal validity of experimental designs are briefly summarized in some books on healthcare research (e.g., Kane, 2005; Keele, 2012; Rubinson & Neutens, 1987; Tappan, 2015), a search of PubMed found they have received minimal attention in medical or other healthcare journals despite the fact that randomized controlled trials (RCTs), which are true experiments, are considered to be the "gold standard" of medical research

(Greenhalgh, 2001; Salmond, 2008). Campbell and Stanley's "Experimental and Quasi-Experimental Designs for Research" is a particularly valuable resource for understanding internal validity and external validity and it can be downloaded from the Internet for free. Explanations of the seven threats to internal validity are given in the following sections.

SEVEN THREATS TO THE INTERNAL VALIDITY OF EXPERIMENTS

History

Campbell's (1957) concept of *history* might be called experience, but *history* in the Campbell conception of threats to internal validity specifically encompasses those things (i.e., specific events) that a study participant experiences during the course of an experiment that are not part of the experiment itself; therefore, they are extraneous variables. In hospital settings, history may include the transfer of patients between units, staff assignment changes, symptom exacerbation, and adverse events associated with treatments. If an experiment takes only a few minutes, *history* is not likely to be a threat to internal validity. Even if an experiment takes a few hours, *history* may not be a threat to internal validity if the study involves hospitalized patients whose experience is limited by the confines of a hospital. However, many experiments may be conducted over, days, weeks, or months. Thus, common everyday experiences like reading or listening to news stories may affect study outcomes. While it may be unlikely that news stories will affect physiological measures of study participants' responses to a medical treatment for cancer, diabetes, or some other medical problem, they are likely to affect psychological outcome measures if the patients learn a new medical treatment is improving health outcomes for these medical ailments or learn that some new treatment has fallen short of its expected efficacy. Participants' outcomes may also be affected by the extent to which they communicate their personal experiences, including perceived treatment effects, with other participants in a study.

Maturation

Whereas *history* involves the experience of external events, *maturation* involves bodily changes. The concept of *maturation* in the context of internal validity encompasses much more than the processes we usually think of as *maturation,* like age-related biological changes. It also encompasses any biological changes that occur with the passage of time, such as becoming hungry, tired, or fatigued, wound healing, recovering from surgery, and disease progression (e.g., stages of cancer or other diseases). Fatigue can occur even during brief interventions with young children, elderly persons, and patients who are feeling ill.

Taylor and Asmundson (2008) provide a hypothetical example of a study in which *maturation* presents a threat to internal validity because the study does not control for it. The example is a "single arm study"[1] of older adults in the early stages of Alzheimer's disease who showed improvement in depression after being treated with a new antidepressant medication. The researchers concluded that the participants' improved mood was due to the medication, when in fact, the depression of many of the participants had remitted naturally as their dementia progressed. As their memory declined, their awareness and insight into their dementia decreased, and they were no longer depressed about the problem.

Testing

The term *testing* is shorthand for any form of measuring study outcomes in study participants. These include physiological measures, which range from having their blood pressure and temperature taken to undergoing an MRI (magnetic resonance imaging). Behavioral tests include such tasks as answering questions on a survey, taking an eye test, or taking a psychological test. Since health researchers often measure a health outcome (the dependent variable) before and after treatment, *testing* becomes a threat to internal validity if the test itself can affect participants' responses when they are tested again. Campbell (1957) called such a test a "reactive measure." This can happen when the test measures participants' knowledge regarding a topic and they learn the correct answers to the questions before taking the test again. It could also happen when a test measures participants' attitudes about a topic and they alter their responses when tested again to give more socially acceptable answers to questions. However, deceit is not the only type of problem posed by repeated testing. The initial test may make some study participants more attuned to the health outcome of interest in a study. An initial measure of personal dietary habits might make a study participant think about his or her poor eating habits and improve them during the course of the study. Similarly, an initial measure of personal exercise might increase a study participant's awareness of his or her lack of exercise and increase how much s/he exercises over the course of the study. These actions would pose threats to the internal validity of experiments designed to improve diet and exercise, respectively, if this participant was in the control group of those experiments; that is, the group of participants who did not receive the intervention.

Measures or tests are less likely to be reactive when they are part of a regular routine, such as measures of temperature, blood pressure, heart rate, and other standard tests patients undergo during their hospital stay. Therefore, the degree to which a researcher can incorporate a measure of the health outcome of interest into the study participants' usual routine, the less likely the measure is to be reactive and, therefore, the less likely it

is to pose a threat to the internal validity of an experiment (Campbell, 1957; Campbell & Stanley, 1966).

Campbell (1957) suggests that a good way to reduce the reactivity of a measure is to embed it in an array of other measures, especially measures that may distract participants from the focus of the study. Unobtrusive observation of patient behavior and reviews of patient charts are nonreactive measures that are very useful if they are applicable to the research question. Webb, Campbell, Schwartz, and Sechrest (1966) discussed the value of observational, archival, and other nonreactive measures in their 1966 book (*Unobtrusive Measures*) and later editions of the book that are still available.

Instrument Decay

Campbell's (1957) concept of *instrument decay* is exemplified by a researcher using a battery-powered device to measure blood pressure in an experiment investigating the effectiveness of a drug to reduce hypertension. Consider an experiment designed for the researcher to measure the blood pressure of the participants before (a "pretest") and after (a "posttest") the participants take the drug for one month. Suppose, however, that unbeknownst to the researcher, the battery has decayed during the month so all the blood-pressure readings taken by the device are lower on the posttest than they were on the pretest.[2] This would severely undermine the internal validity of the experiment. One can imagine a similar experiment to evaluate the effectiveness of a drug to reduce anxiety, in which the researcher assessed anxiety using two battery-powered devices to measure galvanic-skin-response (GSR), a common physiological measure of anxiety. Further imagine that the researcher used one device to measure the GSR of the participants in the experimental group (the participants who took the drug for a month) and the second device to measure the GSR of the participants in the control group (the participants who did not take the drug). If the battery decayed in the device used with experimental participants but not in the device used with control participants, the GSR readings would, hypothetically, be lower for the experimental group than the control group. This would lead the researcher to conclude that the drug reduced anxiety, when, in fact, the observed effect was due to the faulty battery, not the drug.

Campbell and Stanley (1966) changed the term *instrument decay* to *instrumentation* to reflect the fact that any change in measurement ability can pose a threat to internal validity. The term instrument, as used by Campbell (1957) and Campbell and Stanley (1966) is not limited to electronic or mechanical instruments, and applied to any means of measuring the dependent variable, including human researchers or research assistants who observe, judge, rate, and/or otherwise measure a dependent variable. For instance, suppose a new medical resident at an out-patient clinic receives training to observe and rate the functional health-status of patients who

volunteered for a study, using the 0 to 100 Karnofsky Performance Scale (KPS) (e.g., Crooks, Waller, Smith, & Hahn, 1991; Terret, Albrand, Moncenix, & Droz, 2011).[3] After the training, the resident rates all the patients in the study as the pretest, after which half the patients receive a one-month intervention to improve their level of functioning. Because the resident regularly works at the clinic, he becomes better at observing patients and using the KPS by the time he performs the patient ratings for the posttest. This change in *instrumentation* undermines the study's internal validity. A similar problem would arise if someone with extensive experience using the KPS left the research team and had to be replaced with someone new. Since all observers probably get better at observing the more they do it, researcher assistants or other observers on a research team must be thoroughly trained before the start of an experiment.

Statistical Regression

Statistical regression is the tendency for individuals who score extremely high or extremely low, relative to the mean or average, on an initial measure of a variable to score closer to the mean of that variable the next time they are measured on it. *Statistical regression* is more accurately referred to as *regression toward the mean*. It is a threat to internal validity because individuals who are selected for a study because they score high on some measure are likely to score lower on that measure the next time they are tested even without an experimental intervention, whereas individuals who are selected for a study because they score low on some measure are likely to score higher on that measure the next time they are tested even without any experimental intervention (Campbell, 1957; Campbell & Stanley, 1966). Thus, if participants are selected for very low scores on some measure, such as a pretest of reading comprehension, and given an intervention to improve their comprehension of what they read, they are likely to score higher on the posttest even without the intervention to improve their reading comprehension.

An example would be a hospital that wanted to increase its medical interns' awareness of end-of-life issues (e.g., advance directives) and, therefore, might want to offer an in-service. The hospital might decide it would be more efficient to give the in-service only to those interns who seemed to need it the most. As a result, it gave all the interns a test regarding end-of-life issues and selected those who scored the lowest to attend the in-service, and test them after the in-service to see if their scores improved. Since test scores are affected by many things in addition to the knowledge regarding a topic, some interns might have scored very low on the initial test because they were not motivated, they felt tired, they were upset about something that had happened earlier during the day of the test, or were distracted by things that they had to do later that day. Hence, their scores might have increased when they were tested again, even if they had not attended the in-service.

The concept of *statistical regression* or, more precisely, *regression toward the mean* may seem odd or esoteric, but it is quite common, as sports fans should recognize. Although a few teams in any sport may consistently perform better than or worse than the average team, in terms of the number of games they win each year, most teams are always closer to the average, and the teams who do extremely better or worse than average in one year are likely to be closer to the average the next year.

Selection

Selection refers to a potential bias in selecting the participants who will serve in the experimental and control groups; hence it is also known as s*election bias*. The threat to internal validity is that the individuals who are assigned to the experimental and control groups differ from each other in some important ways; that is, that the groups are not equivalent. S*election bias* would be an obvious problem if participants were allowed to choose whether they participated in the control group or the experimental group (i.e., self-selection), but *selection bias* is usually more subtle. For example, a researcher might decide to assign patients at a hemodialysis (HD) center to the experimental and control groups of a study to improve patients' quality-of-life, based on the days of the week they receive treatment [e.g., Monday, Wednesday, and Friday (MWF) vs. Tuesday, Thursday, and Saturday (TTS)]. Although this may seem like a good approach, the patients receiving HD treatment on the MWF schedule probably differ in a number of ways from those receiving it on the TTS schedule. Indeed, the fact that they chose different treatment schedules probably reflects some of these differences; therefore, these differences are confounded with the effects of a quality-of-life intervention, which would undermine the study's internal validity.

Another example of *selection bias* is provided by a hypothetical experiment in which a researcher wants to test the effectiveness of an 8-week family-based, weight-loss program for obese elementary school children. After obtaining permission for the study from the school board and school principals, and distributing flyers regarding the study to parents, the parents begin to enroll their families to participate in the experiment. As the experimental condition is more time-consuming for the researcher to conduct than the control condition is, he decides to "run" the experimental condition (i.e., implement the experimental intervention) first, while he has the time, and run the control condition after the experimental condition is done. Hence, he assigns all the families who immediately enroll in the program to the experimental group, and subsequent families to the control group. Because this decision biases the selection of the participants for the two groups (*selection bias*), the results of the experiment may reflect differences in the families assigned to the experimental and control groups, rather than the effectiveness of the experimental treatment. One such difference might

be that the parents who immediately enrolled their families in the study were more motivated to reduce their child's weight than were the parents who enrolled their families later.

Mortality

Campbell's (1957) *mortality* (sometimes called *experimental mortality*) refers to the differential loss of study participants in the experimental and control groups. This loss is more commonly referred to as the *drop-out rate,* or simply *attrition*. The *drop-out rate* is more likely to be higher for an experimental group than a control group because experimental procedures usually make more demands on and require more commitment and effort from participants. If less committed individuals drop-out of the experimental group, the results may suggest that the intervention is more beneficial than it actually is. For example, individuals who exercise regularly or want to exercise regularly are more likely than other individuals in the experimental group to complete an exercise intervention that lasts several weeks. Thus, results indicating that the experimental intervention was superior to the control condition would be based on a selective subgroup of the original experimental group, and the selective subgroup may be more likely to benefit from the intervention because of their greater motivation to exercise.

INTERNAL VALIDITY OF THREE PRE-EXPERIMENTAL DESIGNS

Campbell described three of what he called pre-experimental designs (Campbell, 1957; Campbell & Stanley, 1966). The three designs are illustrated in Figures 1–3, using the symbols they used to denote the observation or measurement of the dependent variable (**O**) and the presentation of the independent variable (**X**), also referred to as an intervention or treatment.

One-Shot Case Study

Figure 1 shows that this design consists of a single group of participants who receive an intervention (**X**), which is followed by some measure of the dependent variable (**O**). Campbell (1957) said "This design does not merit the title of an experiment" and that he only included it as "a reference point" (p. 298). Fred Kerlinger (1973), who wrote several books regarding research, claimed this design has no scientific value, although he seemed to concede it has clinical value. We think it also has some utility in the context of education,

Group X O

FIGURE 1 One-shot case study.

although it obviously does not control for *history* or *maturation* because it lacks a control group. Hence, there is no way of knowing whether the observed value of the dependent variable is due to the independent variable.

Moreover, the lack of a pretest means there is no way of knowing if the value of the dependent variable was any different after the independent variable was introduced than it was before the independent variable was introduced. The other threats to internal validity do not come into play because of the simplicity of the design.

A study submitted to the *Journal of Health Care Chaplaincy* (*JHCC*) several years ago attempted to use this design to demonstrate that satisfaction with chaplaincy services among the staff of several hospital departments increased after the chaplaincy department made changes to better integrate its services with the services provided by those departments (the study's independent variable). The study was rejected for publication because the authors claimed that staff satisfaction with the chaplaincy department improved after the changes were made without conducting a pretest to measure staff satisfaction with the chaplaincy department before the changes were made; thus, the study could not provide evidence of an improvement in staff satisfaction after the changes were made.

However, *JHCC* has published four articles employing the same design in what are called program description and evaluation studies, which are consistent with the pre-experimental nature of this design. The first article evaluated a curriculum for chaplaincy residents to increase their research skills (Derrickson & Van Hise, 2010), and the last article evaluated a journal club to increase the research skills of chaplaincy residents (Fleenor, Sharma, Hirschmann, & Swarts, 2017). The second one evaluated a curriculum on spiritual assessment for chaplaincy, social work, and medical students (Lennon-Dearing, Florence, Halvorson, & Pollard, 2012), and the third one evaluated a chaplaincy residency program in palliative care (Jackson-Jordan et al., 2017). The four studies had the program participants provide feedback regarding the personal and professional value of the program, the program's implementation (e.g., were program goals clear and were they met), and ways to improve it.

One-Group Pretest-Posttest Design

Figure 2 depicts Campbell's (1957) One-Group Pretest–Posttest Design. The design consists of one group of participants who are given a pretest to measure the dependent variable ($\mathbf{O_{1)}}$, followed by an intervention (**X**), and a posttest to measure the dependent variable again ($\mathbf{O_2}$).

Group O_1 X O_2

FIGURE 2 One-group pretest-posttest design.

If the staff satisfaction study we just mentioned had included a pretest, it would have had a design like the one shown in Figure 2. Similar to the One-Shot Case Study, this design lacks a control group, which prevents the researcher from making the causal inference that the observed value of the dependent variable is due to the independent variable, instead of *history, maturation,* or even *testing*. This design is now known as a "single group study" or a "single arm study," which is the aforementioned design in the hypothetical study described by Taylor and Asmundson (2008), in which *maturation* (disease progression) was confounded with the independent variable (an antidepressant medication), such that the researchers mistakenly concluded that the improved mood of the study participants was due to the antidepressants.

Despite its severe limitations, this design has gained acceptance in certain areas of healthcare research (Ip et al., 2013), especially quality improvement studies (McLaughlin & Kaluzny, 2006; Van Bokhoven, Kok, & Van der Weijden, 2003). However, the Agency for Healthcare Research and Quality of the U.S. Department of Health and Human Services has said it is not appropriate for assessing the effectiveness of clinical interventions (Ip et al., 2013).

If the study participants were selected for a One-Shot Case Study because of their high or low scores, *regression toward the mean* would also pose a threat to internal validity. However, this problem could be remedied by adding a second pretest: $\mathbf{O}_1$ $\mathbf{O}_2$ **X** $\mathbf{O}_3$. The researcher would then compare the values of the dependent variable at $\mathbf{O}_2$ and $\mathbf{O}_3$ to assess the effect of the independent variable **X**, because *regression toward the mean* should have occurred between $\mathbf{O}_1$ and $\mathbf{O}_2$.

JHCC has published three articles that used the design in Figure 2, which is probably the most common design used in program evaluation studies. The three studies provide valuable information about the usefulness of their programs (i.e., their independent variables) without making claims regarding the effectiveness of the independent variables. The first article reported that a workshop on pastoral-care research improved "chaplains" feelings and attitudes regarding research (Murphy & Fitchett, 2009). The second article reported that a chaplain-led spiritual intervention with community-dwelling older adults improved their sense of connection to others and feelings about life, including a sense of gratitude (Grewe, 2017), and the third article reported that a chaplain-led spiritual intervention with U.S. Veterans Administration patients reduced their sense of spiritual distress (Kopacz, Adams, & Searle, 2017).

Static Group Comparison

Campbell (1957) called the third pre-experimental design, the Static Group Comparison, which is shown in Figure 3. Cook and Campbell (1979) later

Group 1	**X**	**O**
Group 2		**O**

FIGURE 3 Static group comparison.

called this a Posttest-Only Design with Nonequivalent Groups. Unlike the first two designs we have discussed, this design consists of two groups: Group 1 receives an intervention before the dependent variable is measured, whereas Group 2 does not receive the intervention.

A researcher might be inclined to use this design to compare the effects of a medical intervention with patients on one hospital unit to the effects of standard care with patients on another hospital unit, but this would be a bad idea, as this design fails to control for *history, maturation,* and *selection bias; selection bias* is always a threat to internal validity unless participants are randomly assigned to the study groups. Nevertheless, this is the basic design used in epidemiological case-control studies, with **X** representing exposure to some disease-promoting agent or event, also known as a risk factor (Kelsey, Thompson, & Evans, 1986; Kleinbaum, Kupper, & Morgenstern, 1982). A common way for epidemiologists to attempt to control for *history* is to match the individuals they select for the unexposed group (the "control," or more appropriately the "comparison" group) with the exposed group as best as they can on different variables, such as profession, type of work, neighborhood, or workplace, depending on the type of exposure being investigated (Kleinbaum et al., 1982). *Maturation* in the most general sense can be achieved by matching for age and gender. Other approaches to selecting a "control" or "comparison" group is to draw random samples from: (a) the general population that appears to be the same as the case population, (b) neighborhoods similar to those where the cases live, and (c) persons seeking medical care at the same hospitals where the cases sought care (Kelsey et al., 1986).

INTERNAL VALIDITY OF THREE EXPERIMENTAL DESIGNS

All true experimental designs entail random assignment of the study's participants to the experimental and control groups at the start of the experiment, which is indicated by **R** in Figures 4–6. Random assignment can be conducted in many ways, including something as simple as flipping a coin if there are only two groups: heads = experimental group, tails = control group, or vice versa.

Posttest-Only Control Group Design

The experimental design Campbell (1957) called the Posttest-Only Control Group Design, which is depicted in Figure 4, is identical to the

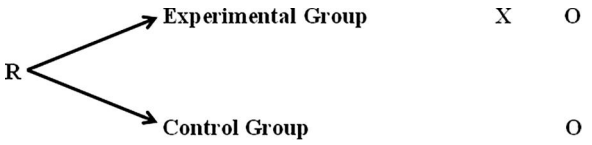

FIGURE 4 Posttest-only control group design.

pre-experimental design that Campbell called the One-Shot Case Study, depicted in Figure 1, except that the Posttest-Only Control Group Design includes the random assignment of participants to groups or conditions (i.e., the experimental condition and control condition). As previously mentioned, *selection bias* is the threat that the participants assigned to the experimental and control groups are not equivalent. Random assignment to groups is used to eliminate the threat of *selection bias* and thereby ensure that the experimental and control groups are equivalent at the start of the experiment.

Although Campbell and Stanley (1966) simply inserted an **R** before each group to indicate random assignment to groups, we think Figure 4 provides a better depiction of random assignment, because it illustrates that the participants are randomly assigned from a common pool of participants who will serve in either the experimental or the control group.

The Posttest-Only Control Group Design is one of the most widely used designs in animal research, especially physiological and behavioral studies on "laboratory" rats and mice. Many animal studies do not require a pretest of the dependent variable because the previous experience of the animals is carefully controlled to make them similar, or more technically, to minimize the variation between the animals: their experience is controlled by the temperature and light/dark cycle of the rooms they live in, the standardized cages they live in, the food they eat, their feeding schedules, and so forth. This greatly reduces the likelihood that they will differ on the dependent variable at the start of an experiment; hence, a pretest of the dependent variable is usually not deemed necessary. Inclusion and exclusion criteria in human medical research also makes it possible to use this design in RCTs because the criteria create a relatively homogeneous pool of participants who are then assigned to the experimental and control groups (Salmond, 2008). This is a worthwhile design for RCTs of the effects of an independent variable on physical/physiological outcomes (i.e., dependent variables), but it is not recommended for studies of psychological outcomes, which usually employ a pretest to measure the dependent variable before the independent variable is introduced.

The remaining threats to the internal validity of such studies, like all studies, are *history* and *maturation,* which are controlled in this design by having a control group (also called the untreated group) that is compared to the experimental group (also called the treated group). The effects of the treatment or intervention are statistically analyzed simply by comparing

the data on the dependent variable from the two groups at point **O**. This analysis can easily be performed with a *t*-test if the data are suitable for such an analysis, in terms of the level of measurement with which the dependent variable is measured [see *JHCC* articles by Jankowski, Flannelly, & Flannelly (2017) regarding *t*-tests, and L. T. Flannelly et al. (2014b) regarding levels of measurement]. A *t*-test is appropriate for analyzing data measured on an interval or ratio scale.

Pretest–Posttest Control Group Design

Because humans vary in many ways, researchers usually want to demonstrate the equivalence of the experimental and control groups before an experiment by conducting a pretest (de Boer, Waterlander, Kuijper, Steenhuis, & Twisk, 2015), as shown in Figure 5. In addition to pretesting the dependent variable, pretests in healthcare studies typically measure the personal characteristics of participants (e.g., socio-demographic variables, general indices of physical mental and health) to see if the experimental and control groups differ in other ways—aside from the dependent variable—that might affect the outcome of the experiment.

This design is the solution to the aforementioned problem posed in the hypothetical study in which the effect of *maturation* (progression of Alzheimer's disease) on depression was mistakenly thought be the effect of antidepressant medication. The control group in this design not only controls for the possible effects of *maturation,* it also controls for the effects of *history, regression toward the mean,* and *instrumentation* on the dependent variable. This design is widely used as a two-arm RCT.

Of course, the experimental and control arms must be conducted during the same period of time, as shown in the diagram. The control arm cannot be conducted after the experimental arm, as it was in our hypothetical example of an experiment on the effectiveness of a family program to reduce childhood obesity. One should think of a large study that is conducted over time as a series of smaller studies consisting of cohorts of experimental and control participants and be aware that threats to internal validity have to be addressed for each cohort. The cohort could be as small as two; that is, whenever two individuals consent to participate in a study, one could be randomly assigned to the experimental group and one to the control group.

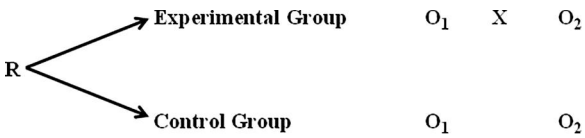

FIGURE 5 Pretest-posttest control group design.

Random assignment to groups controls for *selection bias* in this design, as it did in the Posttest-Only Control Group Design. As *selection bias* is a potential bias when assigning participants to experimental and control groups, the use of inclusion and exclusion criteria can decrease the threat of *selection bias* because they decrease individual differences within the pool of participants on variables that are likely to affect the dependent variable. However, as Campbell (1957) discussed, the homogeneity of the pool of participants that these criteria create reduces the external validity (i.e., the generalizability) of a study's results.

The possible effect of *mortality* can initially be assessed by comparing the *drop-out rates* of the two groups at the end of the study. If the *drop-out rates* are comparable, there is no reason to believe that *mortality* is a threat to the experiment's internal validity. More extensive comparisons of the information collected on the experimental and control participants during the pretest need to be performed if the *drop-out rates* differ substantially between the groups. Although this design controls for *regression toward the mean* because both groups are given a pretest and a posttest, the use of a pretest introduces the threat of a *testing* effect that cannot be isolated with this design.

Campbell recommended statistically analyzing the effects of the independent variable on the dependent variable in this design by calculating difference scores (posttest - pretest) for each participant and comparing the average difference score (also called a change score) of the two groups using a *t*-test. Campbell and Stanley (1966) recommended that the analysis of the difference scores should statistically control for participants' pretest scores by using them as a covariate in the analysis, which requires using a statistical procedure called analysis of covariance (ANCOVA). Although a discussion of ANCOVA lies outside the scope of this article, the analysis recommended by Campbell and Stanley is the preferred way to analyze these data (Takona, 2002) [see Keppel & Wickens (2004), Tabachnick & Fidell, 2013 or a similar book for a description of ANCOVA]. Similar to the *t*-test, ANCOVA is only appropriate for analyzing interval or ratio data.

Naturally, the longer the time-span is between the pretest and the posttest, the greater the likelihood is that *history* and *maturation* will pose a threat to the internal validity of an experiment. Therefore, pretests and posttests should be administered as close to the beginning and end of the intervention as possible. However, a researcher may want to assess the long-term effects of the independent variable on the dependent variable. This can be assessed by adding another posttest ($\mathbf{O_3}$) at some point in time after the initial posttest ($\mathbf{O_2}$) to both groups in Figure 5, which would be shown as $\mathbf{O_2}$ $\mathbf{O_3}$. These data can also be analyzed to statistically control for participants' pretest scores ($\mathbf{O_1}$) using ANCOVA if all the data are interval or ratio variables (Twisk & de Vente, 2008). A statistical procedure called logistic regression is often used to control for pretest

scores when the dependent variable is dichotomous or binomial, such as "case/no-case" (Kleinbaum, Kupper, Nizam, & Rosenberm, 2014).

Solomon Four-Group Design

This design, which is shown in Figure 6, controls for all seven threats to internal validity: *history, maturation, instrumentation, regression toward the mean, selection, mortality,* and *testing*. However, it does not appear to be widely used in healthcare research. This is probably because it is time-consuming and costly to run four groups of participants just to control for the effects of *testing,* especially as: (a) the Pretest-Posttest Control Group Design controls for all of the threats to internal validity except *testing,* and (b) there may be no reason for the researcher to believe the tests used to measure the dependent variable are reactive measures.

Nevertheless, we will explain it briefly. The four groups in the design are created by randomly assigning participants to each group, which controls for *selection bias*. The design is obviously a combination of the designs illustrated in Figures 4 and 5. While Group 1 measures changes due to the independent variable, Group 2 controls for the effects of *history* and *maturation,* which always offer competing explanations for the apparent effect of the independent variable (**X**) on the dependent variable (**O**). Group 2 also controls for the possible confounds of *regression toward the mean* and *instrumentation*.

Since Groups 3 and 4 do not include pretests, they control for the effects of *testing* (the pretest) on the posttest scores of the dependent variable in Groups 1 and 2, respectively. Campbell recommends the 2×2 analysis of variance (ANOVA) shown in Figure 7 to analyze the effects of *testing* on the final (right-hand) measures (**O'**s) of the dependent variable shown in Figure 6. This diagram implies that the analysis assesses the main effect of pretest (Yes or No), the main effect of treatment (**X** = Yes or No) and the interaction of pretest and treatment [see Keppel & Wickens (2004) or a similar book for a description of ANOVA].

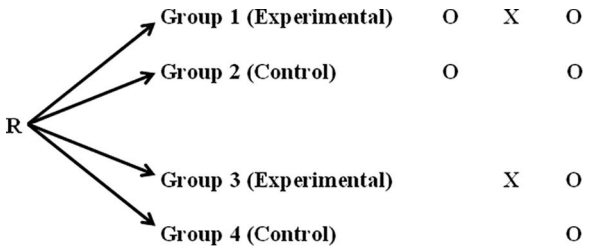

FIGURE 6 Solomon four-group design.

	X = Yes	**X** = No
Pretest = Yes	Group 1	Group 2
Pretest = No	Group 3	Group 4

FIGURE 7 ANOVA to assess the effect of *testing* in the Solomon four-group design.

INTERNAL VALIDITY OF SELECTED QUASI-EXPERIMENTAL DESIGNS

Nonequivalent Control Group Design

Campbell and Stanley (1966) said this was the most frequently used design in educational research and it seems that it still is (Takona, 2002). The design, which is shown in Figure 8, consists of two groups: one that receives an intervention (the independent variable) at point **X** and one that does not. A pretest ($\mathbf{O_1}$) and posttest ($\mathbf{O_2}$) is given to each group. The **A** indicates that the study participants are assigned to each group, but the assignment is not randomized (Campbell & Stanley, 1966). Cook and Campbell (1979), who said it was the most frequently used design in social science research, gave it the unwieldy name, "The Untreated Control Group Design with Pretest and Posttest."

Typically, entire classes of students are assigned to groups in educational research, instead of assigning individual students to groups, which makes the design easy to implement in educational settings. However, the design is subject to *selection bias* because of the lack of random assignment to groups. As an example, students in different classes that are grouped by ability (also known as academic tracking) could cause *selection bias*. Although this would be an obvious problem that a researcher could easily avoid by selecting two classes at the same academic level, the composition of classes might differ for reasons that could be far less obvious to a researcher. As the experimental and control groups are not equivalent because participants are not randomly assigned to groups, the "control" group in this design is usually called a "comparison" group.

This design also applies to the previously described hypothetical study of quality of life of patients undergoing HD, in which the researcher assigned the patients to the experimental and control groups based on the days of the week they attended the HD center (MWF vs. TTS). As previously noted, there may be many reasons why the patients chose to receive HD on MWF

A	Experimental Group	$\mathbf{O_1}$	**X**	$\mathbf{O_2}$
A	Control Group	$\mathbf{O_1}$		$\mathbf{O_2}$

FIGURE 8 Nonequivalent control group design.

vs. TTS, and their reasons could affect the study's outcomes. One reason could be that some patients may choose MWF for convenience because they do not work and they want their weekends free, whereas other patients may choose TTS because they work and they cannot take three weekdays off from their jobs. Another reason might be that relatives and friends drive the patient to and from the HD center, and one of them is only available to do so on weekends.

Like the Pretest–Posttest Control Group Design shown in Figure 5, this design controls for *maturation,* and, to some extent, it controls for *history,* although the experiences of participants may differ somewhat on MWF than TTS. However, we came across a quasi-experimental study comparing the effects of standard care and patient-centered care after hip surgery, which might appear to be a Nonequivalent Control Group Design, but obviously does not control for *history* (Olsson, Karlsson, Berg, Kärrholm, & Hansson, 2014). That study did not control for *history* because the experimental condition (the patients who received patient-centered care) was conducted nine months after the control condition (the patients who received standard care) was conducted, and the control and experimental conditions (or groups) have to be conducted during the same time-period to control for *history*.

An actual Nonequivalent Control Group Design also controls for the effects of *regression toward the mean,* and *instrumentation* on the dependent variable. Once again, potential differences in *drop-out rates* pose the threat of *mortality*. As previously mentioned, however, if the *drop-out rates* are similar, there is no reason to believe that *mortality* is a threat to the study's internal validity. Because this study employs a pretest and a posttest, it is possible that *testing* poses a threat to its internal validity.

Similar to the Pretest–Posttest Control Group Design, the best way to analyze the data is to perform ANCOVA on the posttest–pretest change scores, using the pretest score as a covariate. Although the design does not control for *selection bias,* if the pretest included measures of variables that might be expected to affect the dependent variable, they can also be used as covariates in the ANCOVA. For example, in the study on the quality of life of HD patients, it would be valuable to measure employment and social support during the pretest, and, therefore, they can be controlled for statistically in the analysis of changes in quality of life.

Pretest–Posttest Two Treatment Group Design

Figure 9 illustrates a design, which we named a Pretest-Posttest Two Treatment Group Design. The design has the same flaws and most of the benefits of the Pretest-Posttest Control Group Design, but one cannot be sure the changes observed over time in either group would not have occurred without any treatment. Yet, this design can be useful for comparing different treatment or intervention effects. This is the basic design used by Jankowski,

A	Experimental Group A	$\mathbf{O_1}$	$\mathbf{X_A}$	$\mathbf{O_2}$
A	Experimental Group B	$\mathbf{O_1}$	$\mathbf{X_B}$	$\mathbf{O_2}$

FIGURE 9 Pretest-posttest two treatment group design.

Vanderwerker, Murphy, Montonye, and Ross (2008) to compare changes in pastoral skills, self-reflection, and other variables in clinical pastoral care (CPE) students who took a short/intensive CPE course ($\mathbf{X_A}$) or a long-extended CPE course ($\mathbf{X_B}$).

The length of the intervention (the CPE course) differed, suggesting that the confounding effect of *history* is problematic for the study's internal validity. *Selection bias* is another serious threat to the study's internal validity because the students chose which course they would take. Jankowski et al. (2008) tried to address this threat (a) by obtaining pretest measures of age, gender, years in ministry, years of theological education, and prior CPE training; and (b) using multiple regression to statistically control for the effects of these variables on the change scores of the dependent variables (simple regression is discussed by Flannelly, Flannelly, & Jankowski, 2016).

Cook and Campbell (1979) described a similar design with two treatment groups that they called "The Reversed-Treatment Nonequivalent Control Group Design with Pretest and Posttest." As the name implies, the two independent variables in such a study are designed to produce reverse or opposite effects on the dependent variable. If the opposite directional effects are observed at the posttest, this makes a logically compelling case that the observed effects are due to the independent variables, even though it lacks a control group that does not receive either treatment.

Time-Series Designs

Campbell and Stanley (1966) illustrated several time-series designs, the first of which they simply called **The Time Series Experiment**. This design is particularly useful when there is a system for collecting data on some variable of interest (i.e., the dependent variable) across an extended period of time on a regular periodic schedule. Figure 10 shows a time-series experiment in which the dependent variable is measured each month for 12 months, with the intervention introduced about the sixth month of data collection. The design provides evidence for the effect of the independent variable on the dependent variable, if the introduction of the independent variable (X) alters the dependent variable.

$$\mathbf{O_1 \quad O_2 \quad O_3 \quad O_4 \quad O_5 \quad O_6 \; X \; O_7 \quad O_8 \quad O_9 \quad O_{10} \quad O_{11} \quad O_{12}}$$

FIGURE 10 The time series experiment.

To make the discussion more concrete, assume that a hospital system wanted to conduct a study to evaluate the effectiveness of a pilot program to increase the use of advance directives before investing in a full program. This design would be a reasonable choice because it controls for *testing, regression toward the mean, maturation,* and *mortality*. It controls for *testing* because the measurement of the dependent variable is nonreactive, as it would be the regular monthly summary of the number of patients who signed advance directives. It controls *regression toward the mean* because extreme values in either direction that are due to chance would balance out across the time series. It mainly controls for *maturation* and *mortality* because most of the patients without advance directives are unique patients at each time-point. Campbell and Stanley (1966) said the design controls for *selection bias,* but it fails to control for *instrumentation* and *history*.

The preferred design for this study is shown in Figure 11, which Campbell and Stanley (1966) called The Multiple Time Series Design. A hospital system could easily employ this design, because it could introduce an intervention in one of its hospitals and use another hospital as a control or comparison. If the intervention (**X**) continues over time past $\mathbf{O}_7$, the design would be depicted as a series of **X**'s between $\mathbf{O}_7$ and $\mathbf{O}_{12}$.

The design controls for everything that the design in Figure 10 does and more. Most importantly, it controls for *history*. The most obvious reasons why *history* might pose a threat to the study's internal validity are public efforts to promote the use of advance directives or the recent hospitalization of a family member or friend with a fatal illness who did not have the capacity to make their own health care decisions. It controls for *instrumentation* because the definition of what advanced directives would not be expected to change over time in both hospitals.

There are several ways to analyze both of these designs. With six observations before and after the intervention, one could conduct a single-sample *t*-test on the mean number of pre- and postintervention advance directives in the first design. The analysis of the means should be conducted in the multiple time-series design using ANOVA. It would probably be worthwhile to measure the dependent variable for all the analyses as the percentage of new advance directives by patients who did not initially have them.

Another way to analyze the effect of the independent variable on the dependent variable is to conduct separate regression analyses on the pre- and post-intervention data and statistically compare the slopes of their trend lines. If the dependent variable is measured as a percentage or other

$$\mathbf{O_1\ O_2\ O_3\ O_4\ O_5\ O_6\ X\ O_7\ O_8\ O_9\ O_{10}\ O_{11}\ O_{12}}$$

$$\mathbf{O_1\ O_2\ O_3\ O_4\ O_5\ O_6\ \ O_7\ O_8\ O_9\ O_{10}\ O_{11}\ O_{12}}$$

FIGURE 11 The multiple time series design.

proportion, one could also use the Change-Point Test to assess whether the independent variable affected the dependent variable (Siegel & Castellan, 1988).

Summary

Of the seven threats to the internal validity of an experiment that we discussed, *history* and *maturation* should be thought of as universal threats because they are always present when an independent variable is present. More sophisticated experimental designs tend to introduce new threats to internal validity that must be controlled as they attempt to eliminate other threats. These threats include *testing, instrument decay, regression toward the mean, selection,* and *mortality*. All seven threats to internal validity are problematic because they offer alternate explanations for the observed effects of an experiment on the dependent variable, other than the independent variable. Therefore, it is imperative that researchers understand these threats as they apply to any experimental design chosen and implemented by a researcher. Science moves forward only as far as good research design permits.

Finally, although we only mentioned inclusion and exclusion criteria in passing, we noted that they are useful in health care research to create a relatively homogeneous pool of participants who can be assigned to the experimental and control groups. While the homogeneity of a sample is useful for assessing the experimental effects of an independent variable on a dependent variable, it reduces the external validity (i.e., the generalizability) of the experimental findings (Campbell, 1957).

NOTES

1. The term "single-arm study" is explained later in the article.
2. We do not know if the blood-pressure readings actually decrease when the battery is low, but this is a hypothetical example.
3. The Karnofsky Performance Scale ranges from 0 = Dead to 100 = Normal, no complaints, no evidence of disease. Some other points on the scale include: 20 = Very sick, hospital admission necessary, active supportive treatment necessary; 40 = Disabled, requires special care and assistance; 60 = Requires occasional assistance, but is able to care for most of his/her personal needs; and 80 = Normal activity with effort, some signs or symptoms or disease.

REFERENCES

Bernard, C. (1865/1957). *An introduction to the study of experimental medicine*. New York, NY: Dover.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. doi:10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally & Company.

Cherulnik, P. D. (1983). *Behavioral research: Assessing the validity of research findings in psychology*. Philadelphia, PA: Harper & Row.

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

Crooks, V., Waller, S., Smith, T., & Hahn, T. J. (1991). The use of the Karnofsky Performance Scale in determining outcomes and risk in geriatric outpatients. *Journal of Gerontology*, *46*(4), M139–M144. doi:10.1093/geronj/46.4.m139

de Boer, M. R., Waterlander, W. E., Kuijper, L. D., Steenhuis, I. H., & Twisk, J. W. (2015). Testing for baseline differences in randomized controlled trials: An unhealthy research behavior that is hard to eradicate. *International Journal of Behavioral Nutrition and Physical Activity*, *12*(1), 4. doi:10.1186/s12966-015-0162-z

Derrickson, P., & Van Hise, A. (2010). Curriculum for a spiritual pathway project: Integrating research methodology into pastoral care training. *Journal of Health Care Chaplaincy*, *16*(1–2), 3–12. doi:10.1080/08854720903451030

Flannelly, K. J., Flannelly, L. T., & Jankowski, K. R. B. (2016). Studying associations in health care research. *Journal of Health Care Chaplaincy*, *22*(3), 118–131. doi:10.1080/08854726.2016.1194046

Flannelly, K. J., & Jankowski, K. R. B. (2014). Research designs and making causal inferences from healthcare studies. *Journal of Health Care Chaplaincy*, *20*(1), 23–38. doi:10.1080/08854726.2014.871909

Flannelly, L. T., Flannelly, K. J., & Jankowski, K. R. B. (2014a). Independent, dependent, and other variables in healthcare and chaplaincy research. *Journal of Health Care Chaplaincy*, *20*(4), 161–170. doi:10.1080/08854726.2014.959374

Flannelly, L. T., Flannelly, K. J., & Jankowski, K. R. B. (2014b). Fundamentals of measurement in healthcare research. *Journal of Health Care Chaplaincy*, *20*(2), 75–82. doi:10.1080/08854726.2014.906262

Fleenor, D., Sharma, V., Hirschmann, J., & Swarts, H. (2017). Do journal clubs work? The effectiveness of journal clubs in a clinical pastoral education residency program. *Journal of Health Care Chaplaincy*, 1–14. doi:10.1080/08854726.2017.1383646

Greenhalgh, T. (2001). *How to read a paper: The basics of evidence based medicine*. London, UK: BMJ Books.

Grewe, F. (2017). The Soul's Legacy: A program designed to help prepare senior adults cope with end-of-life existential distress. *Journal of Health Care Chaplaincy*, *23*(1), 1–14. doi:10.1080/08854726.2016.1194063

Ip, S., Paulus, J. K., Balk, E. M., Dahabreh, I. J., Avendano, E. E., & Lau, J (2013). Role of single group studies in Agency for Healthcare Research and Quality comparative effectiveness reviews. Research white paper (Prepared by Tufts Evidence-based Practice Center under Contract No. 290–2007-10055-I) (AHRQ Publication No. 13-EHC007-EF). Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from www.effectivehealthcare.ahrq.gov

Jackson-Jordan, E., Stafford, C., Stratton, S. V., Vilagos, T. T., Janssen Keenan, A., & Greg Hathaway, G. (2017). Evaluation of a chaplain residency program and its partnership with an in-patient palliative care team. *Journal of Health Care Chaplaincy*, 1–10. doi:10.1080/08854726.2017.1324088

Jankowski, K. R., Vanderwerker, L. C., Murphy, K. M., Montonye, M., & Ross, A. M. (2008). Change in pastoral skills, emotional intelligence, self-reflection,

and social desirability across a unit of CPE. *Journal of Health Care Chaplaincy*, *15*(2), 132–148. doi:10.1080/08854720903163304

Jankowski, K. R. B., Flannelly, K. J., & Flannelly, L. T. (2017). The *t*-test: An influential inferential tool in chaplaincy and other healthcare research. *Journal of Health Care Chaplaincy*, 1–10. doi:10.1080/08854726.2017.1335050

Kane, R. L. (Ed.). (2005). *Understanding health care outcomes research* (2nd ed.). Sudbury, MA: Jones & Bartlett.

Keele, R. (2012). *Nursing research and evidence-based practice: Ten steps to success*. Sudbury, MA: Jones & Bartlett Learning.

Kelsey, J., Thompson, W. D., & Evans, A. S. (1986). *Methods in observational epidemiology*. New York, NY: Oxford University Press.

Keppel, G. & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York, NY: Holt, Rinehart, and Winston.

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiolic research: Principles and quantitative methods*. New York, NY: Van Nostrand Reinhold.

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberm E. S. (2014). *Applied regression analysis and other multivariable methods* (5th ed.). Boston MA: Cengage Learning.

Kopacz, M. S., Adams, M. S., & Searle, R. F. (2017). Lectio Divina: A preliminary evaluation of a chaplaincy program. *Journal of Health Care Chaplaincy*, *23*(3), 87–97. doi:10.1080/08854726.2016.1253263

Lennon-Dearing, R., Florence, J. A., Halvorson, H., & Pollard, J. T. (2012). An interprofessional educational approach to teaching spiritual assessment. *Journal of Health Care Chaplaincy*, *18*(3–4), 121–132. doi:10.1080/08854726.2012.720546

McLaughlin, C. P., & Kaluzny, A. D. (2006). *Continuous quality improvement in health care: Theories, implementations, and applications* (3rd ed.). Sudbury MA: Jones and Bartlett.

Mill, J. S. (1859). *A system of logic, ratiocinative and inductive; bring a connected view of the principles of evidence and the methods of scientific investigation*. New York, NY: Harper & Brothers.

Murphy, P. E., & Fitchett, G. (2009). Introducing chaplains to research: "This could help me." *Journal of Health Care Chaplaincy*, *16*(3–4), 79–94. doi:10.1080/08854726.2010.480840

Olsson, L. E., Karlsson, J., Berg, U., Kärrholm, J., & Hansson, E. (2014). Person-centred care compared with standardized care for patients undergoing total hip arthroplasty—A quasi-experimental study. *Journal of Orthopaedic Surgery and Research*, *9*(1), 95. doi:10.1186/s13018-014-0095-2

Rubinson, L., & Neutens, J. J. (1987). *Research techniques for the health sciences*. New York, NY: Macmillan Publishing.

Salmond, S. S. (2008). Randomized controlled trials: Methodological concepts and critique. *Orthopaedic Nursing*, *27*(2), 116–122. doi:10.1097/01.nor.0000315626.44137.94

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York, NY: McGraw-Hill.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Takona, J. P. (2002). *Educational research: Principles and practice*. Lincoln, NE: Writers Club Press.

Tappan, R. M. (2015). *Advanced nursing research* (2nd ed.) Sudbury, MA: Jones & Bartlett Learning.

Taylor, S., & Asmundson, G. J. G. (2008). Internal and external validity in clinical research. In D. McKay (Ed.), *Handbook of research methods in abnormal and clinical psychology* (pp. 23–34). Los Angeles, CA: Sage.

Terret, C., Albrand, G., Moncenix, G., & Droz, J. P. (2011). Karnofsky Performance Scale (KPS) or Physical Performance Test (PPT)? That is the question. *Critical Reviews in Oncology/Hematology*, *77*(2), 142–147. doi:10.1016/j.critrevonc.2010.01.015

Twisk, J. W., & de Vente, W. (2008). The analysis of randomised controlled trial data with more than one follow-up measurement. A comparison between different approaches. *European Journal of Epidemiology*, *23*(10), 655–660. doi:10.1007/s10654-008-9279-6

Van Bokhoven, M. A., Kok, G., & Van der Weijden, T. (2003). Designing a quality improvement intervention: A systematic approach. *Quality and Safety in Health Care*, *12*(3), 215–220. doi:10.1136/qhc.12.3.215

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago, IL: Rand McNally.