The Principle of Parallelism in the Design of Studies to Estimate Treatment Effects

Charles S. Reichardt
University of Denver

An effect is a function of a cause as well as of 4 other factors: recipient, setting, time, and outcome variable. The principle of parallelism states that if a design option exists for any 1 of these 4 factors, a parallel option exists for each of the others. For example, effects are often estimated by drawing a comparison across recipients who receive different treatments. The principle of parallelism implies that an effect can also be estimated by drawing a comparison across settings, times, or outcome variables. Typologies of methodological options are derived from the principle of parallelism. The typologies can help researchers recognize a broader set of options than they would otherwise and thereby improve the quality of research designs.

Keywords: research design, quasi-experiment, randomized experiment, causality, treatment effect

Estimating the effects of treatments or interventions is a mainstay of research in psychology as well as in the other behavioral and social sciences. For example, estimating the effects of interventions is one of the primary means of testing theories. Estimating the effects of treatments is also one of the primary means of evaluating efforts to ameliorate practical problems in health, education, and welfare. That estimating treatment effects plays a central role in psychological research is also evidenced by the prominence of methodological treatises on experimental and quasi-experimental designs, such as by Boruch (1997), Campbell and Stanley (1966), Cook and Campbell (1979), Cronbach (1982), Judd and Kenny (1981), Mohr (1995), Riecken et al. (1974), and Shadish, Cook, and Campbell (2002). The present article introduces a rule, called the principle of parallelism, to assist researchers when crafting designs to estimate treatment effects.

The principle of parallelism derives from the fact that the size of a treatment effect can depend not only on the nature of the treatment but also on four other factors: the recipient, setting, time, and outcome variable. The principle of parallelism states that if a design option exists for any one of the four factors of recipient, setting, time, and outcome variable, then a parallel option exists for each of the other factors. For example, as is well known, effects often are estimated by drawing a comparison across recipients who receive different treatments. What is not as well known, but is implied by the principle of parallelism, is that an effect can also be estimated by drawing a comparison across either settings, times, or outcome variables that are assigned different treatments. Understanding the principle of parallelism can help researchers recognize a broader range of design options than they would otherwise. And the more options that are recognized, the more likely it is that the most effective option can be chosen. In what follows, I explicate the principle of parallelism by showing how it operates in a variety of ways that are central to the task of estimating effects. In so doing, I derive several typologies of methodological options and show how they can be used to improve research designs. The typologies involve (a) comparisons for estimating treatment effects, (b) threats to validity, and (c) methods for ruling out threats to validity.

The Size of an Effect

Understanding the principle of parallelism requires understanding what is meant by the effect of a difference between two treatments. Consider two points in time, where Time 1 is the time at which treatments are implemented and Time 2 is a later time at which the effects of the treatments are assessed. The effect of a difference between Treatments A and B is defined as the difference between what would have happened at Time 2 if Treatment A had been implemented at Time 1 and what would have happened at Time 2 if Treatment B had been implemented at Time 1.
2 if Treatment B, instead of Treatment A, had been imple-
mented at Time 1 but everything else at Time 1 had been the
same. For example, suppose you had taken two aspirins an
hour ago because you had a headache and you no longer
have a headache now. Further, suppose that if you had taken
two placebo pills an hour ago instead of the two aspirins but
everything else had been the same including your headache,
you would still have the headache now. Then relief from
your headache is an effect of the difference between having
taken the aspirins and having taken the placebos. The pre-
ceeding definition of an effect is in accord with the defini-
tions of effects given by others such as Boruch (1997),
and West, Biesanz, and Pitts (2000).
A special case of the preceding definition of an effect of
the difference between two treatments arises when Treat-
ment B is equal to the absence of Treatment A. In that case,
it is meaningful to talk about the effect of Treatment A without
reference to Treatment B. Because it is easier to talk
about the effect of a treatment than about the effect of a
difference between two treatments, the present article is
phrased in terms of the effect of a treatment rather than the
effect of a difference between two treatments. This simpli-
fication results in no loss in generality because everything
said in the present article applies equally to both cases.
However, in practice, Treatment B is seldom, if ever, equal
to the mere absence of Treatment A, so the special case
seldom, if ever, arises (Holland, 1986). And even if one of
the treatments is equal to the absence of the other, research-
ers should be aware that an effect is always caused (and
defined, as I have done above) by the difference between
two treatments.
As can be seen from the preceding definition of an effect,
the size of a treatment effect depends on the treatment. The
size of a treatment effect can also depend on four other
factors: the recipient, setting, time, and outcome variable, as
is described next.
The recipient is the entity or entities (usually people or
other species) that receive the treatment and on which
outcomes are measured. A recipient can be either an indi-
vidual or a group of individuals (such as a school or com-
unity). The size of an effect can depend on the recipients
as well as the treatment because a treatment can have
different effects on different recipients. For example, aspirin
can reduce headache pain in most individuals but might
increase headache pain in those allergic to aspirin.
The setting is the environment in which the treatment is
implemented and assessed. A setting might be described by
terms of both physical features (e.g., laboratory, voting
booth, subway, highway, or sidewalk) and functional char-
acteristics (e.g., conversation, job interview, or child’s play
date). The size of a treatment effect can vary across settings.
For example, aspirin might have a large effect on headache
pain in a quiet seaside cottage but only a small effect in a
noisy bar.
Time refers to the chronological times at which both a
treatment is implemented and an outcome variable is as-
essed, which thereby specifies the lag between these two
times. The size of a treatment effect can differ both for
different times in history and for different time lags. For
example, the effect that occurs a minute after taking aspirin
can be different than the effect that occurs an hour after
taking aspirin.
The size of an effect could be assessed on any of a wide
variety of outcome variables. For example, the effect of
taking aspirin could be assessed on the dimensions of head-
ache pain, arthritis pain, stomach upset, or susceptibility to
heart attack, to name just a few. The size of a treatment
effect can vary across different outcome variables. For
example, aspirin can have a substantial effect on variables
such as arthritis pain and susceptibility to heart attack but
very little or no effect on variables such as visual acuity and
compulsivity.
Because the size of an effect can vary with the treatment,
recipient, setting, time, and outcome variable, I will call
these five factors the size-of-effect factors. Conceptually,
the five size-of-effect factors are the who (recipient), how
(treatment), what (outcome variable), where (setting), and
when (time) of the size of an effect. Recognition of these
cfive factors, in one fashion or another, is common place in
the methodological literature, including in the works by
Shadish et al. (2002), Judd and Kenny (1981), and Cronbach
(1982), though the factor of time is sometimes given rela-
tively short shrift, as I note later.
The principle of parallelism states that if a methodologi-
cal option exists for any one of the four size-of-effect
factors of the recipient, setting, time, and outcome variable,
a parallel option exists for each of the other three factors as
well. Both a strong and a weak form of the principle of
parallelism can be distinguished. In the strong form, paral-
lelism holds for the treatment factor as well as for the other
four size-of-effect factors of the recipient, setting, time, and
outcome variable. In the weak form, the principle of paral-
lelism holds for only the four size-of-effect factors of the
recipient, setting, time, and outcome variable. The size-of-
effect factor of the treatment is excluded in the weak form
of the principle of parallelism because the treatment plays a
unique role in estimating effects: The treatment is the only
size-of-effect factor that varies across the treatment condi-
tions in the comparison that defines a treatment effect. Both
the strong and weak forms of the principle of parallelism are
illustrated below. The weak form applies to typologies that
categorize comparisons used to estimate effects (including
comparisons used to rule out threats to validity). The strong
form of the principle of parallelism applies everywhere else,
including typologies of threats to validity. The distinction
between strong and weak forms is not of primary impor-
tance but is drawn simply to acknowledge that sometimes the size-of-effect factor of the treatment is included in the principle of parallelism and sometimes not.

A Typology of Comparisons for Estimating a Treatment Effect

The comparison, described in the preceding section, that defines the effect of a difference between two treatments is called the ideal comparison (Reichardt & Mark, 1998). The ideal comparison is used to define what a treatment effect means but is impossible to obtain in practice because it requires that everything, other than the treatment, remains the same at Time 1. In any comparison that can be drawn in practice, something else will necessarily vary at Time 1 along with the treatment. The something else that will necessarily vary along with the treatment is either the recipient, setting, time, or outcome variable. Indeed, all four of these size-of-effect factors tend to vary with the treatment in practical comparisons, though variation in one of the these four size-of-effect factors is more prominent (usually by design) than in the others. The size-of-effect factor that varies most prominently with the treatment in a comparison will be called the prominent size-of-effect factor. The recipient, setting, time, and outcome variable can each be the prominent size-of-effect factor in a comparison. In this way, four types of comparisons can be distinguished on the basis of which of the four size-of-effect factors is most prominent. Each of these four types of comparisons is described below. That four types of comparisons (recipient, time, setting, and outcome variable comparisons) can be distinguished is an instance of the (weak form of the) principle of parallelism. The benefits of distinguishing among these four types of comparisons are described in a subsequent section.

**Recipient Comparisons**

To assess the effects of the source of a message on agreement with the contents of the message, a (hypothetical) researcher created two copies of a historical speech. In one copy, the speech is attributed to Thomas Jefferson, whereas in the other, the speech is attributed to Karl Marx. The researcher randomly assigned these different forms of the speech to recipients in a study, and each recipient indicated his or her degree of agreement with the contents. The effect of the source of the message on agreement was assessed by comparing the responses across the two treatment conditions.

Comparisons such as in the preceding example will be called recipient comparisons. Because treatment groups are composed of different recipients, the recipients vary with the treatment. In other words, in comparing the outcomes from the two treatment conditions, the researcher is comparing outcomes from different recipients.

The recipient is seldom the only size-of-effect factor that varies with the treatment in a recipient comparison. If the recipients are studied one at a time, for example, time will vary along with recipients, and the setting will vary across the treatment conditions, along with recipients and times. That is, if different recipients are studied at different times, the conditions of the setting vary as well, even if only slightly. Even if all the recipients are studied at the same time and in the same room, the setting varies with the treatment because, for any two recipients, the setting for Recipient X contains Recipient Y, whereas the setting for Recipient Y contains Recipient X.

But although more than one size-of-effect factor can vary across the treatment conditions, a comparison is a recipient comparison when the recipient is the size-of-effect factor that varies most prominently with the treatment. This is the case in the preceding example concerning the effect of the source of a message on agreement with that message. In the preceding example, the recipients (e.g., persons) are also both the replicates in the statistical analysis and the units of analysis.

**Time Comparisons**

Do the cyclamates contained in diet beverages increase the severity of a clinical patient’s migraine headaches? To answer this question, a (hypothetical) researcher created 60 otherwise indistinguishable doses of beverage, 30 of which contain cyclamates and 30 of which do not. The patient consumed 1 of the 60 doses of beverages each morning for 60 days. Which day each type of beverage is consumed is determined at random. Each afternoon, the patient recorded the severity of any headache. The effect of cyclamates on headache pain is estimated by comparing the responses across the two treatment conditions.

Comparisons such as in the preceding example will be called time comparisons. Because the two treatment conditions are administered on different days, time varies with the treatment. In other words, in comparing the outcomes from the two treatment conditions, the researcher is comparing results at different times.

Of course, as time varies in a time comparison so can the recipient and the setting. For example, many of an individual’s characteristics (such as mood, attention, and fatigue) can change over time. Similarly, many aspects of a setting can change over time. But although more than one size-of-effect factor can vary across the treatment conditions, a comparison is a time comparison when time is the size-of-effect factor that varies most prominently with the treatment. This is the case in the preceding example concerning the effect of cyclamates on headaches. In the preceding example, times (e.g., days) are also both the replicates in the statistical analysis and the units of analysis.
Setting Comparisons

To estimate the effect that the presence a police officer has on the likelihood that pedestrians jaywalk at intersections, a (hypothetical) researcher randomly assigned street corners either to have or not have a police officer present. The effect of the treatment is estimated by comparing, across the two treatment conditions, the proportion of pedestrians who jaywalk at each of the street corners.

Comparisons such as in the preceding example will be called setting comparisons. Because the two treatment conditions (police officer vs. no police officer) are administered in different settings (i.e., at different street corners), settings vary with the treatment. In other words, in comparing the outcomes from the two treatment conditions, the researcher is comparing results from different settings.

Other size-of-effect factors besides the setting can vary with the treatments in setting comparisons. For example, different recipients (i.e., pedestrians) may walk past the different street corners, so recipients can also vary with the treatments. But although more than one size-of-effect factor can vary across the treatment conditions, a comparison is a setting comparison when the setting is the size-of-effect factor that varies most prominently with the treatment. This is the case in the preceding example concerning the effect of police officers on jaywalking. In the preceding example, settings (e.g., street corners) are also both the replicates in the statistical analysis and the units of analysis.

Outcome Variable Comparisons

To assess how well a cartoon format can teach a child with learning disabilities to pronounce the letters of the alphabet, a (hypothetical) researcher selected half the letters of the alphabet at random and presented them in a cartoon video. The video was shown to the child, and his or her ability to pronounce each of the letters in the alphabet was measured. The effect of the video was assessed by comparing the child’s responses on the letters that were taught to the responses on the letters that were not taught on the video.

Comparisons such as in the preceding example will be called outcome variable comparisons. Because the ability to pronounce each letter of the alphabet is a different outcome variable and because responses to the letters taught and not taught are being compared, the outcome variable varies with the treatment. In other words, in comparing the outcomes from the two treatment conditions, the researcher is comparing results from different outcome variables.

Within-subject designs used in laboratory psychological research, where different stimuli are used to implement different treatment conditions, are usually outcome variable comparisons. For example, in studies of the effects of massed versus distributed practice, a group of recipients viewed a list of words in which the words were randomly assigned to be repeated in either massed or distributed fashion (Campbell & Stanley, 1966; Underwood, 1966). The recipients were then asked to recall the words in the list. The effect of massed versus distributed practice was assessed by comparing the responses with the two sets of words. In this case, the recall of each word is a different outcome variable.

Other factors besides the outcome variables can also vary along with the treatments in outcome variable comparisons. For example, responses to different letters cannot be assessed at exactly the same time for a child, so time can vary along with the treatment and outcome variable. Similarly, once time varies, characteristics of both the recipient and the setting can also vary even if only slightly. For example, as time varies, a recipient can change because of experience or fatigue. But although more than one size-of-effect factor can vary across the treatment conditions, a comparison is an outcome variable comparison when the outcome variable is the size-of-effect factor that varies most prominently with the treatment. This is the case in the preceding example concerning the effect of the cartoon format on learning letters. In the preceding example, the outcome variables (e.g., letters) are also both the replicates in the statistical analysis and the units of analysis.

Adding the Distinction Between Randomized and Nonrandomized Experiments to the Typology of Comparisons

In the preceding section, four types of comparisons were distinguished on the basis of which of the four size-of-effect factors of recipient, setting, time, and outcome variable varied most prominently with the treatment. The prominent size-of-effect factor in each of these four types of comparisons can vary either randomly or nonrandomly with the treatment conditions. If the prominent size-of-effect factor varies randomly with the treatment conditions, the comparison is called a randomized experiment. If the prominent size-of-effect factor varies nonrandomly with the treatment, the comparison is called a nonrandomized experiment. In addition, two types of nonrandom assignment will be distinguished: those based on an explicit quantitative ordering and those not based on an explicit quantitative ordering. The four types of comparisons and the three types of treatment assignment (one random and two nonrandom types of assignment) are crossed. That is, each of the four types of comparisons can be implemented using each of the three types of treatment assignment. The complete set of possibilities is depicted in the typology in Table 1.
ment varied randomly with the prominent size-of-effect factor. That is, treatment conditions were assigned randomly to either (a) the recipients (e.g., individuals in the study of the effects of the source of a message on agreement with the message), (b) times (e.g., days in the study of cyclamates), (c) settings (e.g., street corners in the study of jaywalking), or (d) outcome variables (e.g., letters of the alphabet in the study of the cartoon video). As a result, these comparisons are all randomized experiments and make up the first column in Table 1.

Of course, varying a treatment randomly across the prominent size-of-effect factor does not guarantee that the treatment varies randomly across any of the other size-of-effect factors. For example, it is possible to assign recipients to treatment conditions at random but assess the recipients on different days and in different settings, neither of which vary randomly with the treatments. Because of the benefits of random assignment, researchers often try either to (a) randomly assign the prominent size-of-effect factor in a way that also makes other size-of-effect factors vary randomly with the treatment conditions or (b) introduce randomization repeatedly so that both the prominent and nonprominent size-of-effect factors vary randomly across the treatment conditions. For example, randomly assigning participants to treatment conditions and assessing the recipients from one treatment condition altogether on 1 day and the recipients from the comparison condition altogether on a different day would not be an effective way to distribute the times of participation randomly across treatment conditions. But scheduling individual times for recipients to participate in a study and then assigning the recipients to the different treatment conditions on the basis of a coin flip at the time they arrive in the laboratory would distribute both recipients and times of participation randomly across the treatment conditions.

Nonrandom Assignment Based on an Explicit Quantitative Ordering

Rather than varying randomly with treatment conditions, the prominent size-of-effect factor could vary nonrandomly. A special type of nonrandom assignment arises when the units of the prominent size-of-effect factor (e.g., individuals, days, street corners, or letters) are quantitatively ordered on an observed characteristic and assigned to treatment conditions based on that ordering. An example of such assignment is where the units with scores below a cutoff value on the quantitative ordering are assigned to one treatment condition, whereas the units with scores above the cutoff value are assigned to the alternative treatment condition (see Rubin, 1974, for probabilistic assignment rather than assignment based on a strict cutoff score). Comparisons where treatment conditions are assigned nonrandomly on the basis of an explicit quantitative ordering of the units of the prominent size-of-effect factor make up the middle column in Table 1.

When the recipient is the prominent size-of-effect factor with, say, individuals as the units and with the individuals ordered and assigned to a compensatory treatment using a cutoff score on a quantitatively assessed measure of need, the comparison is a regression-discontinuity design (Shadish et al., 2002). The regression-discontinuity label arises because of the way the treatment effect is estimated. Conceptually, the outcome variable is regressed onto the quantitative assignment variable separately in each treatment group (Reichardt, Trochim, & Cappelleri, 1995; Trochim, 1984). When a treatment has a constant effect across levels of the quantitative assignment variable, that effect is manifested by a break or discontinuity in the two regression lines at the cutoff score.

When time is the prominent size-of-effect factor with, for example, days as the units and with days before a cutoff date
being assigned to one treatment condition and days after that date being assigned to an alternative treatment condition, the comparison is an interrupted time-series design (Shadish et al., 2002). The interrupted time-series label arises because of the way the effect of the treatment is estimated. Conceptually, the outcome variable is regressed onto the variable representing chronological time separately in each treatment group (McCleary & Hay, 1980). When a treatment has an immediate and abrupt effect, that effect is manifested by a break or interruption in the two regression lines at the cutoff time.

When the setting is the prominent size-of-effect factor with, for example, street corners as the units, street corners could be ordered quantitatively, say, according to location or traffic volume (cf. Reichardt & Cook, 1979) and that quantitative ordering could be used to determine assignment to treatment conditions. For example, to cut down on jaywalking, police officers could be assigned to street corners on the basis of a quantitative assessment of pedestrian traffic, with those street corners exhibiting the greatest volume of pedestrians being assigned police officers. Or to reduce accidents, traffic lights could be added to intersections on the basis of a quantitative assessment of the number of accidents that were reported at the intersections during a preceding period of time, with lights being added to those intersections having the most accidents. Assigning street corners or intersections to different treatment conditions using a cutoff score on a quantitative variable would result in a discontinuity across settings design. An estimate of the treatment effect would be derived in a fashion parallel to the way a treatment effect is estimated in regression-discontinuity and interrupted time-series designs (cf. Marcantonio & Cook, 1994).

Finally, when the outcome variable is the prominent size-of-effect factor with, say, letters as the units, letters could be quantitatively ordered and assigned to treatment conditions using a cutoff score. For example, letters could be ordered on the basis of their position in the alphabet or on the basis of their frequency of appearance in literature. The resulting comparison would be a discontinuity across outcome variables design. Again, an estimate of the treatment effect would be derived in a fashion parallel to the way a treatment effect is estimated in regression-discontinuity and interrupted time-series designs.

Perhaps it is worth noting that, compared with randomized experiments, the class of discontinuity designs described above often require more complex statistical analyses. For example, the literature on the analysis of data from interrupted time-series designs has long recognized that autocorrelations often exist among time-ordered observations, which can bias the results if not taken into account in the statistical analyses (McCleary & Hay, 1980). Though it is not as commonly discussed, similar forms of dependence among the ordered observations can also arise in discontinuity-type designs where recipients, settings, and outcome variables are the prominent size-of-effect factor.

**Nonrandom Assignment Not Based on an Explicit Quantitative Ordering**

In a nonequivalent recipient design, recipients are the prominent size-of-effect factor and are assigned to treatment conditions neither randomly nor on the basis of an explicit quantitative ordering. For example, a nonequivalent recipient design would result if recipients self-selected themselves into treatment conditions based on nonrandom and unmeasured personal characteristics.

In parallel fashion, when the prominent size-of-effect factor is either times, settings, or outcome variables, each can vary with the treatment conditions nonrandomly without an explicit quantitative ordering. These comparisons are nonequivalent time, nonequivalent setting, and nonequivalent outcome variable designs, respectively, when times, settings, and outcome variables are the prominent size-of-effect factors. These types of comparisons make up the third column in Table 1.

Estimating a treatment effect using a nonequivalent comparison requires that the researcher take account of the effects of initial differences between the units in the different treatment conditions. Most of the literature concerned with statistical methods for taking account of the effects of initial differences focuses on taking account of initial differences between recipients in the context of nonequivalent recipient comparisons (e.g., Heckman & Robb, 1985; Reichardt, 1979; Rosenbaum, 1984; Rosenbaum & Rubin, 1983; Winship & Morgan, 1999), but much the same methods could be used if the units were settings, times, or outcome variables.

**The Most Commonly Used Comparisons**

That each of the four types of comparisons (i.e., the four rows in Table 1) can be implemented with any of the three types of assignment to treatment conditions (i.e., the three columns in Table 1) is another instance of the (weak form of the) principle of parallelism. That is, any of the options in Table 1 that are available for one of the four size-of-effect factors are available for each of the others as well.

The comparisons in some cells in Table 1 are used in practice and discussed in texts on research methods more frequently than the comparisons in other cells. In addition, different fields of study tend to use some types of designs more often than other types. For example, laboratory research in social psychology uses randomized recipient comparisons more often than other types of comparisons. Research in cognitive psychology most often uses randomized recipient comparisons and randomized outcome variable comparisons (under the headings of between-subjects designs and within-subjects or repeated-measures designs, re-
spectively). Interrupted time-series designs (i.e., nonrandomized time comparisons based on an explicit quantitative ordering) are widely used, for example, in research on behavioral modification. In my experience, program evaluation in the social sciences uses types of recipient comparisons in the following order of frequency: nonrandomized and not based on an explicit quantitative ordering, randomized, and nonrandomized but based on an explicit quantitative ordering. Setting comparisons are frequently used in epidemiology (especially nonrandomized setting comparisons not based on an explicit quantitative ordering), agricultural research (especially randomized setting comparisons), and community psychology.

Among randomized comparisons (i.e., among the cells in the first column of Table 1), recipient comparisons and outcome variable comparisons are often discussed in statistics texts in the social sciences under the heading of between-subjects designs and within-subjects or repeated-measures designs, respectively. Campbell and Stanley (1966) devoted separate sections to designs where times and outcome variables are assigned to treatment conditions at random (Designs 8 and 9, respectively) but considered them to be quasi-experiments rather than randomized experiments. Emphasis on these two design possibilities seems to have largely disappeared from the subsequent literature on quasi-experimentation.

Among nonrandom comparisons based on an explicit quantitative ordering (i.e., among the cells in the second column of Table 1), time comparisons (e.g., interrupted time-series designs) and recipient comparisons (e.g., regression-discontinuity designs) are the only ones used or discussed with any frequency. Among nonrandomized comparisons not based on an explicit quantitative ordering (i.e., among the cells in the third column of Table 1), recipient and setting comparisons are the most commonly prescribed in the quasi-experimental literature, though designs that incorporate nonrandomized outcome variable comparisons as adjuncts to other types of comparisons (e.g., designs with nonequivalent dependent variables; see Shadish et al., 2002) are also discussed.

In the absence of a well-developed theory of the fit between types of comparisons and research circumstances (cf. Shadish & Cook, 1999), it makes sense, when planning a study, to consider as complete a range of design options as possible. Table 1 lays out a greater variety of design options than is typically recognized by practitioners or emphasized in any one methodological work. For example, because of the designs and examples that are emphasized, I suspect most readers of widely used treatises on quasi-experimentation (such as Judd & Kenny, 1981; Mohr, 1995; Shadish et al., 2002) come away with the sense that designs in only some of the cells in Table 1 are possible. The typology in Table 1 may enable researchers to identify potentially useful designs that might otherwise be overlooked. For example, outcome variable comparisons (both randomized and nonrandomized but based on an explicit quantitative ordering) could be of value in evaluations of educational programs but are seldom used, perhaps partly because they are seldom recognized in methodological writings in applied social research. Consider the summative evaluation of the effects of Sesame Street during its 1st year of production, which used a relatively weak nonequivalent recipient design (Ball & Bogatz, 1970; Cook et al., 1975). With a little forethought, the program designers could easily have implemented a rigorous randomized outcome variable design by choosing, at random, which letters of the alphabet to teach and which letters not to teach on the show during its inaugural year.

Of course, the three types of assignment (the three columns in Table 1) are generally not equally effective in ruling out threats to validity. Random assignment tends to be most effective, and nonrandom assignment based on an explicit quantitative ordering tends to be the next most effective (Shadish et al., 2002). However, it is not always possible to implement the most effective types of assignment. When designing a study, I consider each of the possibilities in Table 1, but I tend to give preference to those further to the left.

A Typology of Threats to Validity

The principle of parallelism leads to a novel typology of threats to validity, which is the topic introduced next. Many typologies of threats to validity have been proposed previously, including ones by Cronbach (1982), Judd and Kenny (1981), and Kruglanski and Kroy (1975; see Mark, 1986). The typology that has evolved over time in Campbell (1957), Campbell and Stanley (1963, 1966), Cook and Campbell (1975, 1979), and Shadish et al. (2002) has been the most influential and widely used. The principle of parallelism provides the typology proposed herein with a simple and unifying structure that is not present in any of these other typologies. Benefits of this structure are described after my typology is presented.

After a comparison in Table 1 is implemented, an estimate of the size of the treatment effect is derived. When reporting the results, the estimate of the size of an effect is

---

1 One might wonder whether the traditional labels of between-subjects designs and within-subjects designs would be better than the novel ones I am using. The drawback to the traditional labels is that they do not distinguish among all four types of comparisons in my presentation. For example, the between-subjects label could refer to either recipient or setting comparisons, so the label is inadequate for distinguishing between these two types of designs. Similarly, the within-subjects label could apply to both and therefore does not distinguish between outcome variable and time comparisons.
packaged in what will be called a size-of-effect statement. A size-of-effect statement has two parts: a label that specifies the effect that is purportedly being estimated and a numerical estimate for the size of that effect. The numerical estimate will be designated by E. Everything else in a size-of-effect statement is the label. A size-of-effect statement has the following general form: the size of the effect of treatment, Tx, for recipient, R, in setting, S, at time, T, on outcome variable, V, is equal to estimate, E. In shorthand notation, a size-of-effect statement is as follows: The size of the effect for Tx, R, S, T, and V equals E.

It is possible that a size-of-effect statement is incorrect (or, what means the same thing, a size-of-effect statement is invalid). A size-of-effect statement is invalid when the numerical estimate and the label given to that estimate do not match. In other words, invalidity arises when the estimate that is derived from a comparison is not given a correct label in a size-of-effect statement. For example, a researcher might estimate an effect using men as the recipients but conclude the estimate applies to both men and women. If the effect of the treatment differs for men and women, the size-of-effect statement is invalid because the estimate that was produced (i.e., the estimate of the effect for men) does not match the label that was given to that estimate (i.e., the effect for men and women). An example comes from Kohlberg’s (1964) research on moral development, which was conducted using adolescent boys alone. The results became a staple in developmental textbooks where they were (at least implicitly) attributed to both genders, until Gilligan (1982) argued that Kohlberg’s interpretations did not hold for adolescent girls because, among other things, adolescent boys and girls differ in how they differentially value the rights of individuals versus the preservation of relationships.

That an estimate does not match the label specified in a size-of-effect statement means that the estimate matches an alternative label. For example, in the preceding illustration, men is the alternative label for the recipients that should have been used in place of the label men and women. An alternative label is said to provide an alternative explanation or rival hypothesis for the results. The source of an alternative label is called a threat to validity (Campbell, 1957; Campbell & Stanley, 1966).

As noted above, a size-of-effect statement is invalid if its two parts (estimate and label) do not match. The five size-of-effect factors (treatment, recipient, setting, time, and outcome variable) comprise the label in a size-of-effect statement, and invalidity can result because of a mismatch between the estimate and the label of any of these five factors. Each type of mismatch is described below. What results is a typology of threats to validity. That each of the five size-of-effect factors can be the source of a mismatch (i.e., the source of a threat to validity) is an instance of the (strong form of the) principle of parallelism.

Mismatch of the Treatment

In a recipient comparison, different groups of recipients are compared. Differences between the recipients in the treatment groups are called selection differences (Campbell & Stanley, 1966). Selection differences can cause a difference in the observed outcomes. When they do, a conclusion that the outcome difference is due solely to the effect of the treatment would be invalid. The invalidity is the result of a mismatch of the treatment, that is, a mismatch between the true cause of the estimate and the label that is given to the cause (i.e., the treatment component of the label). In particular, the estimate is caused by both the treatment and selection differences, but only the treatment is included in the label. In Shadish et al.’s (2002) nomenclature, a mismatch of the treatment such as due to selection differences is a threat to internal validity. All threats to internal validity are mismatches of the treatment.

A mismatch of the treatment could also arise because of what Cook and Campbell (1979) called threats to the construct validity of the cause. For example, an observed outcome difference could be due to the effect that the active ingredients in a medication have on the recipients as well as to the effects that the experimenter’s expectancies about the medication have on the recipients. In this case, a mismatch of the treatment would arise if the conclusion were drawn that an observed outcome difference was due solely to the effect of the active ingredients (such as aspirin) and not also to the effect of the experimenter’s expectancies. More is said about the distinction between internal and construct validity below.

Mismatch of the Recipient

Mismatches of the recipient arise, among other ways, when estimates of treatment effects are said to apply to a more general class of recipients than the class to which they do apply. For example, Schachter (1982) found claims in the clinical literature that treatments for quitting smoking have relapse rates of between 75% and 90%. However, Schachter noted that these results come from studies only of

---

2 This is a simplification because a complete size-of-effect statement is more complex. In particular, rather than specifying that a labeled size of effect equals a numerical estimate, a complete size-of-effect statement specifies that a labeled size of effect is contained within a range of estimates with a certain degree of probability (Reichardt & Gollob, 1987). However, this additional complexity is irrelevant for present purposes and so will be ignored.

3 By taking validity and truth to be equivalent, I am defining validity slightly differently than how it is defined in the Campbellian tradition (cf. Shadish et al., 2002), but the difference is irrelevant for present purposes. The presentation uses my definition because it is simpler.
those who seek treatment because they have been unable or unwilling to quit smoking by themselves:

Our view of the intractability of the addictive states has been molded largely by the self-selected, hard-core group of people who, unable or unwilling to help themselves, go to therapists for help, thereby becoming the only easily available subjects for studies of recidivism and addiction. (p. 437)

It is likely that the relapse rate for smoking-cessation treatments is lower for those who can quit smoking by themselves than for those unable to quit by themselves. If so, the estimates of 75%–90% relapse rates were labeled invalidly because of a mismatch of the recipient; that is, the results apply only to those who seek professional assistance to quit smoking, but the results were said to apply to all smokers. In the nomenclature of Cook and Campbell (1979), mismatches of the recipient are threats to external validity.

Mismatch of the Setting

Mismatches of the setting arise, among other ways, when estimates of treatment effects are said to apply to a more general class of settings than the class to which they do apply. For example, the results of studies of the effects of psychotherapy on children in the laboratory have been said to apply to clinical settings as well. But it has been argued that effects may be substantially smaller in settings outside the laboratory, such as mental health clinics (Weisz, Weiss, & Donenberg, 1992; though see Shadish, Matt, Navarro, & Phillips, 2000; Shadish et al., 1997). If this criticism is correct, the conclusion that results from laboratory studies apply to clinical settings as well would be invalid because of a mismatch of the setting. Similarly, it has been suggested that work training programs are most effective under conditions of economic prosperity. Therefore, a researcher who estimates the effects of a work training program in an environment of economic prosperity but concludes the results hold under more general economic conditions is drawing an invalid conclusion because of a mismatch of the setting.

Mismatches of the setting can also arise if the effect of a treatment depends on how many other recipients receive the treatment. For example, an SAT preparation course might increase the chances an individual gets accepted to college, but not if all college applicants take the course. One of the purposes of Rubin’s (1980, 1990; West et al., 2000) stable-unit-treatment-value assumption (SUTVA) is to avoid mismatches where the effect of a treatment on one recipient depends on the treatment received by others. In the nomenclature of Cook and Campbell (1979), mismatches of the setting are threats to external validity.

Mismatch of the Time

Estimates of treatment effects are often said to apply to a greater variety of time lags than the time lags to which they do apply. Such mismatches of the time most commonly occur when temporal variation in the effectiveness of the treatment is not recognized. For example, deeming a social intervention ineffective implies the treatment has little or no effect at all reasonable time lags. But the treatment might have been assessed only at time lags much too short for its effects to be realized. Similarly, antibiotics and pesticides lose effectiveness over time because bacteria and insects develop resistances, and advertising campaigns lose their legs, as is said in the trade. But the effects of these treatments are often described as if they are constant over time. Both Cook and Campbell (1979) and Shadish et al. (2002) discussed threats to validity due to mismatches of historical time but not (as far as I can tell) due to mismatches of time lags. In the nomenclature of Cook and Campbell (1979), mismatches of historical time are threats to external validity.

Mismatch of the Outcome Variable

No measure is a measure of a single theoretical construct alone. For example, tests used to assess individual differences in mathematical ability also contain variance due to individual differences in, say, reading ability, among many other constructs. Similarly, treatments never influence single constructs. For example, a treatment intended to improve mathematical ability might also influence reading ability. Concluding that an effect on one construct (say, mathematical ability) has been estimated when the estimate is at least partly due to the effect on another construct (say, reading ability) produces a mismatch of the outcome variable; that is, the outcome variable to which a result applies is mismatched with the outcome variable to which the result is said to apply.

For example, making decisions in a group, as opposed to individually, was originally characterized as causing a shift toward riskiness (Bem, Wallach, & Kogan, 1965; Wallach & Kogan, 1965), but the effect has been alternatively interpreted as causing a shift toward social desirability rather than riskiness (Abelson, 1995). If the alternative interpretation is correct, the original interpretation is invalid because of a mismatch of the outcome variable; the outcome variable was originally labeled riskiness when what had actually been observed was an effect along the dimension of social desirability. In the nomenclature of Cook and Campbell (1979), mismatches of the outcome variables are threats to the construct validity of the effect.

A Typology of Mismatches of the Treatment

(Including a Typology of the Different Sources of Threats to Internal Validity)

As explained in the preceding section, a threat to validity results from a mismatch between the label and the estimate of an effect. The preceding section described five types of
mismatches (or threats to validity): one type for each of the five size-of-effect factors. The present section concerns just one of these five types of threats to validity: mismatches of the treatment. Five sources of mismatches of the treatment can be distinguished. Perhaps not surprisingly, the five sources of mismatches are the five size-of-effect factors. The present section explains how each of the five size-of-effect factors can be the source of a mismatch of the treatment.

Taken together, the preceding and present sections produce a two-level typology of threats to validity, as displayed in Table 2. The five classes of threats to validity that are numbered 1–5 in Table 2 were described in the preceding section. The present section describes the five classes of threats to validity labeled A–E in Table 2, which are the five sources of mismatches of the treatment.

The present section is divided into subsections, where separate subsections describe each of the five sources of a mismatch of the treatment that are labeled A–E in Table 2. The first subsection is entitled Differences in the Treatments. Threats to validity that fall into this category are, in the nomenclature of Cook and Campbell (1979), threats to the construct validity of the cause. The next four subsections concern differences in the other four size-of-effect factors of recipient, setting, time, and outcome variable. Threats to validity that fall in these categories are, in the nomenclature of Cook and Campbell (1979), threats to internal validity. Two implications are worth noting (and are discussed in further detail later). First, the typology that is being proposed points out an important commonality as well as the essential difference between threats to the construct validity and threats to internal validity.

Table 2

<table>
<thead>
<tr>
<th>Category and size-of-effect factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mismatch of the treatment</td>
<td></td>
</tr>
<tr>
<td>A. Differences in the treatments</td>
<td>Threats to construct validity</td>
</tr>
<tr>
<td>B. Differences in the recipients</td>
<td>Threats to internal validity</td>
</tr>
<tr>
<td>C. Differences in the settings</td>
<td>Threats to internal validity</td>
</tr>
<tr>
<td>D. Differences in the times</td>
<td>Threats to internal validity</td>
</tr>
<tr>
<td>E. Differences in the outcome variables</td>
<td>Threats to internal validity</td>
</tr>
<tr>
<td>2. Mismatch of the recipient</td>
<td></td>
</tr>
<tr>
<td>3. Mismatch of the setting</td>
<td></td>
</tr>
<tr>
<td>4. Mismatch of the time</td>
<td></td>
</tr>
<tr>
<td>5. Mismatch of the outcome variable</td>
<td></td>
</tr>
</tbody>
</table>

Note. A threat to validity is a mismatch between the label and the estimate of a treatment effect. The table partitions threats to validity into five categories (labeled 1–5) on the basis of which size-of-effect factor (treatment, recipient, setting, time, or outcome variable) is the source of the mismatch. In addition, mismatches of the treatment are partitioned into five categories (labeled A–E) on the basis of whether the mismatch is caused by the differences in the treatments, recipients, settings, times, or outcome variables.
of self-perception instead (Bem, 1972). A substance might cause cancer only when ingested in large quantity, but the statement is erroneously made that cancer results from ingesting even small amounts (McGinley, 1997). Or a researcher might mistakenly conclude that a treatment has no effect when it was implemented with insufficient strength or integrity to detect its otherwise substantial effect (Sechrest, West, Phillips, Redner, & Yeaton, 1979). Mismatches of the treatment that are due to mislabeled differences in the treatments are, in the nomenclature of Cook and Campbell (1979), threats to the construct validity of the cause.

**Differences in the Recipients**

As noted above, in recipient comparisons, differences between the recipients in the treatment conditions (i.e., selection differences) can produce a mismatch of the treatment if outcome differences are attributed to the treatment alone. The presence of selection differences is obvious in nonrandomized recipient comparisons, but selection differences arise even when recipients are randomly assigned to treatments. Random assignment makes selection differences random and therefore zero in expected value. But in any given research instance, random selection differences are likely to be nonzero, in which case their effects are confounded with the treatment effect. In randomized recipient comparisons, the purpose of statistical significance tests and confidence intervals is to take account of the confound due to random selection differences (Reichardt & Gollob, 1987, 1989; Reichardt & Mark, 1998).

Confounds due to differences in recipients can arise in recipient comparisons as well as in comparisons where recipients are not the prominent size-of-effect factor. For example, with time comparisons, differences in recipients can be confounded with the treatment when there are multiple recipients, some of whom are lost from the study over time. In the nomenclature of Shadish et al. (2002), the preceding confound is a threat to internal validity due to attrition or experimental mortality. Conversely, the pool of recipients could increase in size over time, which is not a threat to validity listed in Shadish et al. (2002) but which Mark, Reichardt, and Sanna (2000) labeled augmentation. But even without either attrition or augmentation, recipients can be confounded with the treatment in time comparisons because characteristics (such as the mood or fatigue) of the recipients can change over time. For example, confounds in time comparisons that arise because recipients grow older or fatigued are called threats to internal validity due to maturation (Shadish et al., 2002).

**Differences in the Settings**

In setting comparisons, settings are necessarily confounded with the treatment and therefore are a potential source of threats to internal validity, even when settings are randomly assigned to treatments. Confounds due to differences in settings can also arise when settings are not the prominent size-of-effect factor. In time comparisons, for example, settings are confounded with the treatment because the setting at one point in time is never identical to the setting at later times. Threats to internal validity due to history and local history (Shadish et al., 2002) are differences in settings that are confounded with the treatment.

**Differences in the Times**

In time comparisons, time is necessarily confounded with the treatment and therefore is a potential threat to internal validity. For example, in an interrupted time-series design, seasonal differences can be confounded with the treatment. Confounds with time can also arise when time is not the prominent size-of-effect factor. For example, differences in times arose in a nonrandomized recipient comparison when blood from one treatment group was assessed later in time (after being left out in the air longer) than the blood from a comparison group (Dooley, 1995; Snyder, 1974). This difference in the time lags between treatment implementation and outcome assessment produced a mislabeling of the treatment because part of the effect ascribed to the treatment was actually due to differences caused by different lengths of exposure to the air. Threats to validity due to differences in time lags do not appear to be mentioned in Shadish et al. (2002) but fall under the rubric of internal validity.

**Differences in the Outcome Variables**

In outcome variable comparisons, outcome variables are necessarily confounded with the treatment and therefore are a potential threat to internal validity, even when the outcome variables vary randomly across the treatment conditions. Confounds due to differences in outcome variables can also arise when the outcome variable is not the prominent size-of-effect factor. Differences in outcome variables are a threat to internal validity that Shadish et al. (2002) labeled instrumentation. Observer expectancies can be a source of instrumentation differences. For example, observer expectancies can differentially alter the way observations are recorded in the treatment groups so the treatment appears to have the effect the observers expect.

**The Utility of the Proposed Typology of Threats to Validity**

The typology of threats to validity that is presented in Table 2 is meant to complement alternative typologies of threats to validity, such as the Campbellian ones (e.g., Cook & Campbell, 1979; Shadish et al., 2002), not to replace them. One of the primary uses of a typology of threats to validity is to help researchers identify the threats that are most plausible when either designing or critiquing a study.
The simple structure of my typology, which follows directly from the principle of parallelism, can help researchers recognize plausible threats when used in conjunction with other typologies. For example, consider the long list of threats to internal validity that have been usefully documented in the Campbellian tradition. My typology categorizes the sources of threats to internal validity in a simple way that can help researchers recognize them without having to memorize the Campbellian shopping list. In concrete terms, when assessing a study for threats to internal validity, researchers should think through the size-of-effect factors (see Categories A–E in Table 2) and ask how differences in each might be confounded (i.e., vary along) with the treatment (which would include both confounds involving the prominent factor in a comparison and the incidental confounds that arise from the factors that are not the most prominent one). In this way, being mindful of the principle of parallelism can help researchers recognize threats to validity that might otherwise be overlooked. In this regard, it is perhaps worth noting that the extensive lists of threats to internal validity in Campbell (1957), Campbell and Stanley (1963, 1966), Cook and Campbell (1975, 1979), and Shadish et al. (2002) have all overlooked two threats to internal validity (i.e., augmentation and unequal time lags) that are easily envisioned from the perspective of my typology.

Note also that the typology I am proposing puts the size-of-effect factor of time on equal footing with the other size-of-effect factors. In other typologies, such as the one by Shadish et al. (2002), this is not the case. For example, the list of threats to external validity in Shadish et al. (2002, p. 87) includes interactions with the four size-of-effect factors of units (i.e., recipients), treatments, outcomes, and settings but omits interactions with time, which is a significant oversight from the perspective of my typology.

In some typologies (such as in Cronbach, 1982), the size-of-effect factor of time is subsumed under the rubric of setting and therefore not considered separately. But the separation of setting and time can be justified from the perspective of the principle of parallelism in several ways, including the typology of designs in Table 1. For example, to subsume time under the rubric of setting would be to subsume the interrupted time-series design under the rubric of the interrupted settings design. But the difference between the interrupted time-series design and the interrupted settings design is no less than the difference between, say, the interrupted time-series design and the regression-discontinuity design, and drawing a distinction between the latter two designs has long been a staple in the literature on quasi-experimentation. In addition, it is important to recognize that the size-of-effect factor of time includes the specification of the length of the time lag between treatment and effect and not just the specification of a historical point in time. Subsuming the size-of-effect factor of time under the rubric of setting is likely to gloss over time lags, and there is a growing recognition in the literature that concern for time lags should play a larger, rather than smaller, role in the conceptualization and implementation of both design and analysis (Cole & Maxwell, 2003; Gollob & Reichardt, 1987, 1991; MacCallum & Austin, 2000; Reichardt, 2002; Reichardt & Gollob, 1986). There is a justifiably long-standing tradition in western thought of distinguishing not just who, how, what, and where but also when. That tradition is honored by the principle of parallelism.

In addition, because it is derived from the principle of parallelism, my typology reveals commonalities (and differences) in threats to validity that may not be as obvious from the Campbellian typology. For example, the underlying logic by which researchers take account of threats to validity is the same for all types of threats to validity (Reichardt, 2000). This conclusion has important practical consequences and is a direct consequence of my typology and the principle of parallelism but is not as easily recognized from the Campbellian perspective. Or for another example, consider the distinction between internal validity and the construct validity of the cause. According to Shadish et al. (2002), the difference between these two types of validity is that threats to internal validity are confounds that “could have occurred in the absence of the treatment,” whereas threats to construct validity “would not have occurred had a treatment not been introduced” (p. 95). But this distinction is inadequate because there are many instances where threats to internal validity, such as those that are due to differential attrition, instrumentation, testing, and differential history, would not have occurred if the treatment had not been present. My typology clarifies the line of demarcation between these two types of validity. Both threats to internal validity and threats to the construct validity of the cause are mismatches of the treatment (i.e., mismatches between the size-of-effect estimate and the label of the treatment for that estimate). The difference lies in the sources of the mismatch, as denoted by Categories A–E of the typology in Table 2. When the mismatch of the treatment is conceptualized as due to a difference in the treatments, it is a threat to the construct validity of the cause. When the mismatch of the treatment is conceptualized as due to a difference in recipients, settings, times, or outcome variables, it is a threat to internal validity.

Ruling Out Threats to Validity

The principle of parallelism is relevant not just to the task of recognizing threats to validity but also to ruling them out. The complete set of strategies for ruling out threats to validity is described in Reichardt (2000). One of the strategies is called elaboration and involves two estimates. When a threat to validity is present, it means the original estimate (\(ESTIMATE_1\)) is equal to the treatment effect...
(TREATMENT EFFECT) plus the effect of the threat to validity (EFFECT OF THREAT):

\[ \text{ESTIMATE}_1 = \text{TREATMENT EFFECT} + \text{EFFECT OF THREAT}. \]

For example, when a threat to validity due to selection differences is present, \( \text{ESTIMATE}_1 \) equals the effect of the treatment (TREATMENT EFFECT) plus the effect of selection differences (EFFECT OF THREAT). In elaboration, the effect of the treatment and the effect of the threat to validity, which are confounded in the first estimate (\( \text{ESTIMATE}_1 \)), are disentangled by adding a second estimate (\( \text{ESTIMATE}_2 \)). Elaboration can take a variety of forms, which are distinguished by the nature of the second estimate (Reichardt, 2000). In the method of elaboration called vary the size of the treatment effect, the effect of the threat to validity is the same in \( \text{ESTIMATE}_1 \) and \( \text{ESTIMATE}_2 \), but the size of the effect of the treatment varies. In notational form, the two estimates are as follows:

\[ \begin{align*}
\text{ESTIMATE}_1 &= \text{TREATMENT EFFECT} + \text{EFFECT OF THREAT} \\
\text{ESTIMATE}_2 &= (A \times \text{TREATMENT EFFECT}) + \text{EFFECT OF THREAT},
\end{align*} \]

where \( A \) denotes the degree to which the sizes of the treatment effects differ in the two estimates. A treatment effect is evidenced (above and beyond the effect of the threat to validity) by finding that the two estimates are not equal. An example is provided by Leibowitz and Kim (1992; Azar, 1994), who assessed the effect of a naturally occurring substance (galanin) on weight gain. Injecting galanin into rats increased their weight compared with noninjected controls (\( \text{ESTIMATE}_1 \)). But trauma from the injection was a threat to validity. In a second estimate (\( \text{ESTIMATE}_2 \)), rats injected with a substance that blocks the natural production of galanin lost weight compared with noninjected controls. These results ruled out trauma as a threat to validity because the method uses \( \text{ESTIMATE}_2 \) to estimate and then subtract the effect of the threat to validity. Many design features that rule out threats to validity operate via the estimate-and-subtract method. For example, a control time series of observations is used to rule out the threat of history in an interrupted time-series design via the estimate-and-subtract method. The experimental time series, derived from recipients who receive the treatment, produces \( \text{ESTIMATE}_1 \), which reflects the effects of both the treatment and history, whereas the control time series, derived from recipients who do not receive the treatment but who are susceptible to the same history effects, produces \( \text{ESTIMATE}_2 \). The difference between \( \text{ESTIMATE}_1 \) and \( \text{ESTIMATE}_2 \) is an estimate of the treatment effect free from the effects of history.

Although the examples used above to describe the two methods of elaboration both involve threats to internal validity (i.e., mismatches of the treatment), the strategies of elaboration (as well as all strategies for ruling out threats to validity) are equally applicable to mismatches of any of the five size-of-effect factors (Reichardt, 2000). This is another instance of the strong form of the principle of parallelism.

The principle of parallelism is relevant to the strategy of elaboration in another way as well. In elaboration, the first and second estimates come from different sources. In accord with the (weak form of the) principle of parallelism, the different sources can be either different recipients, settings, times, or outcome variables. Examples of how the first and second estimates in elaboration can be derived from each of these different sources are given below.

In the ensuing examples, the designs that are used to rule out threats to validity via elaboration are creative. I often find that to be the case for the strategy of elaboration: That is, cleverness is often involved in coming up with the second estimate that is used to disentangle the confounded effects of a threat to validity from the effects of the treatment. Such design creativity can be fostered by recognizing all of the four sources of differences between the first and second estimates, as illustrated below and as implied by the principle of parallelism. In other words, to improve their creativity in ruling out threats to validity using the method of elaboration, researchers should think through all the different ways of obtaining two different estimates to disentangle treatment and threat; namely, by using different recipients, settings, times, or outcome variables.

**Different Recipients**

Wagenaar (1981, 1986) used an interrupted time-series design to assess the effects of an increase in the legal age of drinking (from 18 to 21 years of age) on traffic injuries. The important point for the present discussion is that ESTI-
MATE\textsubscript{1} and ESTIMATE\textsubscript{2} came from different recipients. ESTIMATE\textsubscript{1} was derived from a time series of injuries to 18–20-year-olds, which evidenced effects of the intervention as well as effects of history. ESTIMATE\textsubscript{2} was derived from a time series of injuries to 21–24-year-olds, which was free from the effects of the intervention but shared the effects of history. ESTIMATE\textsubscript{1} revealed a reduction in fatalities, whereas ESTIMATE\textsubscript{2} showed no difference, which ruled out the threat to validity via the method of elaboration.

**Different Settings**

Anderson (1989) examined the effects of a law mandating the use of seat belts by those riding in the front seats of automobiles. The important point for the present discussion is that ESTIMATE\textsubscript{1} and ESTIMATE\textsubscript{2} came from different settings. ESTIMATE\textsubscript{1} was derived from a time series of injuries to riders in the front seat of cars, which evidenced effects of the intervention as well as history effects. ESTIMATE\textsubscript{2} was derived from a time series of injuries to riders in the back seat of cars, which should have had either no effect or a reduced effect of the intervention but shared the same effects of history. ESTIMATE\textsubscript{1} revealed a reduction in injuries, whereas ESTIMATE\textsubscript{2} showed no difference in injuries, which ruled out the threat to validity via the method of elaboration.

**Different Times**

Ross, Campbell, and Glass (1970) examined the effects of a crackdown on drunken driving in Great Britain where, it is important to note, most drinking takes place in pubs. ESTIMATE\textsubscript{1} and ESTIMATE\textsubscript{2} came from different times. ESTIMATE\textsubscript{1} was derived from a time series of traffic casualties that occurred during the hours pubs were open. This time series evidenced the effects of the intervention as well as the effects of history. ESTIMATE\textsubscript{2} was derived from a time series of traffic casualties that occurred during the hours pubs were closed. This time series was free from the effects of the intervention but shared the effects of history. ESTIMATE\textsubscript{1} revealed a reduction in casualties, whereas ESTIMATE\textsubscript{2} showed no difference in casualties, which ruled out the threat to validity via the method of elaboration.

**Different Outcome Variables**

Braucht et al. (1995) examined the effects of case management on alcohol abuse. ESTIMATE\textsubscript{1} and ESTIMATE\textsubscript{2} came from different outcome variables. ESTIMATE\textsubscript{1} was derived from the outcome variable of alcohol use and was obtained from a nonrandom comparison of individuals receiving different amounts of case-management services that were designed to reduce substance abuse. This comparison evidenced effects of the intervention as well as the effects of selection differences in motivation, which might have caused recipients to self-select different amounts of the services. ESTIMATE\textsubscript{2} was derived from a comparison of the same groups of individuals on the outcome variable of quality of social relationships, which should have been affected by differences in motivation but not very much by the intervention. ESTIMATE\textsubscript{1} revealed a reduction in alcohol use, whereas ESTIMATE\textsubscript{2} showed no difference in the quality of social relationships, which ruled out the threat to validity via the method of elaboration.

**Summary and Conclusions**

The size of a treatment effect is a function of the treatment as well as of the recipient, setting, time, and outcome variable. The five factors of treatment, recipient, setting, time, and outcome variable, are called size-of-effect factors. The principle of parallelism states that, in the design of a study to estimate a treatment effect, parallel methodological options exist for the four size-of-effect factors of recipient, setting, time, and outcome variable. In other words, for any design in which one of these four size-of-effect factors plays a given role, parallel designs are possible where each of the other three size-of-effect factors plays that role instead.

Both a strong and a weak form of the principle of parallelism can be distinguished. In the strong form, parallelism holds for the treatment factor as well as for the other four size-of-effect factors of the recipient, setting, time, and outcome variable. In the weak form, the principle of parallelism holds only for the four size-of-effect factors of the recipient, setting, time, and outcome variable.

**Types of Comparisons**

Estimating the size of an effect requires drawing a comparison between what happened when a treatment was implemented and what happened when an alternative treatment was implemented. In the ideal comparison for estimating a treatment effect, the treatment alone varies across the comparison conditions, but in practice, at least one of the other size-of-effect factors varies along with the treatment. In accord with the (weak form of the) principle of parallelism, four types of comparisons can be distinguished, based on which of the four size-of-effect factors varies most prominently with the treatment. Also in accord with the (weak form of the) principle of parallelism, the most prominent size-of-effect factor can be assigned to treatment conditions in one of three ways: randomly, nonrandomly based on an explicit quantitative ordering, or nonrandomly not based on an explicit quantitative ordering. Crossing the four types of comparisons with the three types of assignment rules creates the $4 \times 3$ typology of designs given in Table 1.
The prominent size-of-effect factor is often the only source of within-condition replicates in a study (cf. Judd & Kenny, 1981). In elaborate designs, however, replicates can derive from more than one source including (in parallel fashion) recipients, settings, times, and outcome variables. The notion of replicates has important implications for drawing generalizations. Statistical generalizations are permitted only across the replicates in the study and, in particular, only across replicates that are randomly sampled. Recognizing which of the size-of-effect factors is a source of replicates can help researchers appreciate how far their statistical generalizations both do and do not reach. Different replicates answer different questions about the generality of the results.

**Size-of-Effect Statement**

A treatment comparison is implemented to derive an estimate of the effect. The estimate of an effect is reported in a size-of-effect statement, which consists of the estimate of the size of the effect and a label for that estimate. As per the (strong form of the) principle of parallelism, a complete label includes a specification of all five of the size-of-effect factors.

The size-of-effect statements that appear in the literature seldom, if ever, include specification of all five of the size-of-effect factors. Often the omission of some of the size-of-effect factors is of little consequence because the context of the study makes appropriate specification of the missing factors obvious. However, sometimes the omission of a size-of-effect factor is critically important and evidences a fundamental flaw in the design. For example, it is commonplace to use structural equation modeling to estimate causal effects with cross-sectional data (wherein both the cause and the effect are measured at the same point in time) without appreciating the importance of the time lag that must exist between a cause and its effect. Because an effect never arises instantaneously with a cause, it is of no value to estimate causal effects that have zero time lags. But because the size of an effect inevitably varies over time, researchers need to specify which nonzero time lag, among all the possibilities, is the one being estimated. But researchers who use structural equation modeling with cross-sectional data seldom, if ever, make clear the time lag at which they intend to assess the effect, much less whether the data they have collected at a single point in time permit them to assess the effect at the desired time lag (Cole & Maxwell, 2003; Gollob & Reichardt, 1987, 1991; MacCallum & Austin, 2000; Reichardt, 2002; Reichardt & Gollob, 1986). Recognition of the principle of parallelism might help increase awareness of the critical role that time lags play in causal modeling and lead to stronger designs.

Another advantage of thinking in terms of complete size-of-effect statements is that it cautions researchers against drawing conclusions that are overly general, while at the same time encouraging them to consider the relevance of generalizations along the dimensions of each of the five size-of-effect factors. A size-of-effect statement is but a single point on a highly multifaceted response surface of sizes of effects that is a function of innumerable diverse variations in treatments, recipients, settings, times, and outcome variables. Perhaps researchers have little choice but to assume, in the absence of further knowledge, that gradients along this response surface are relatively flat. But if they do so, researchers should be aware that the slopes of the gradients at most locations on the response surface are in fact largely unknown and therefore that any presumption of flatness should be held only very tentatively. Extrapolation of causal effects beyond the range of assessed treatments, recipients, settings, times, and outcome variables can be particularly risky (Cook, 1993; Shadish et al., 2002).

**Size Versus Direction or Existence of an Effect**

Let the question of whether an effect is zero or nonzero be called the question of the existence of an effect. Let the question of whether an effect is positive, negative, or zero be called the question of the direction of an effect. Finally, let the question of the numerical size of an effect be called the question of the size of an effect. Then knowledge of the direction of an effect entails knowledge of the existence of an effect, but knowledge of existence does not necessarily entail knowledge of direction. Similarly, knowledge of the size of an effect necessarily entails knowledge of both existence and direction but not vice versa. The implication is that direction is more informative than existence, and size is more informative than both direction and existence.

Although knowing the size of an effect is more informative than knowing only the direction or existence of an effect, attention in the methodological literature is often focused on direction or existence rather than on size (Reichardt & Gollob, 1987). For example, Campbell and Stanley’s (1966) focus was on the existence of an effect when they stated that internal validity asks the question “Did in fact the experimental treatments make a difference in this specific experimental instance?” (p. 5). A similar focus can be found in both Cook and Campbell (1979) and Shadish et al. (2002). In contrast, a focus on size would mean asking instead “How much, if any, of a difference did the experimental treatment make in this specific experimental instance?” A focus on the direction or existence of an effect rather than on its size is also common in empirical research. For example, a focus on direction or existence rather than size is evidenced by the widespread use of statistical significance tests instead of confidence intervals (Abelson, 1995; Harris, 1997; Reichardt & Gollob, 1997).

As its name suggests, a size-of-effect statement is explicitly concerned with the size, and not just the existence or
direction, of an effect. That a size-of-effect statement focuses on size is in large part a derivative of the principle of parallelism. If researchers are to explicate the logic for estimating an effect in the most general and encompassing terms, they should focus on estimating the size and not just the direction or existence of effects. Focusing on size-of-effect statements (at least conceptually, if not in practice) would be one way to foster a focus on size.

**Threats to Validity and Other Criticisms**

A size-of-effect statement is invalid if the estimate of an effect mismatches the label given to that estimate. In accord with the (strong form of the) principle of parallelism, each of the five size-of-effect factors in a label can be the source of a mismatch. In addition, a mismatch of one of the size-of-effect factors (the treatment) can arise because of a difference between the treatment conditions in any of the five size-of-effect factors (which is another instance of the strong form of the principle of parallelism). In this way, the principle of parallelism provides the basis for the two-level typology of threats to validity given in Table 2. Because it characterizes the sources of threats to validity and is easy to remember, the typology in Table 2 can help researchers recognize threats to validity that might otherwise be overlooked. The typology can also help methodologists understand commonalities (as well as differences) among types of threats to validity. For example, the lower (embedded) level of the typology in Table 2 reveals that both threats to internal validity and threats to the construct validity of the cause are mismatches of the treatment but differ in the size-of-effect factor that is the source of the mismatch.

In the present article, a threat to validity was defined as a mismatch in an explicit statement (in this case, a size-of-effect statement). One advantage of tying the definition of validity to an explicit statement is that it allows other types of criticisms to be more clearly distinguished. A threat to validity is the criticism that a size-of-effect statement is incorrect. Besides being incorrect, a size-of-effect statement could also be criticized, for example, for being either imprecise or irrelevant. The Campbellian typologies tend to include criticisms of imprecision and irrelevance under the rubric of validity, rather than treating these criticisms separately. For example, threats to statistical conclusion validity in Shadish et al. (2002) address the criticism of imprecision and not just the criticism of incorrectness. The principle of parallelism generalizes to other criticisms besides the criticism of invalidity, but such generalizations will be best constructed when the nature of the other criticisms are clearly defined and distinguished, which includes defining them with reference to explicit statements, such as size-of-effect statements.

**Ruling Out Threats to Validity**

In accord with the principle of parallelism, the logic by which threats to validity are ruled out is the same whether the threat to validity is due to a mismatch of the treatment, recipient, setting, time, or outcome variable (Reichardt, 2000). In addition, different methods for ruling out threats to validity can be implemented using a variety of design options that abide by the principle of parallelism. For example, in the method of elaboration, the effect of a treatment is disentangled from the effect of a threat to validity by combining two estimates. In accord with the (weak form of the) principle of parallelism, the two estimates used in elaboration can be derived from either different recipients, settings, times, or outcome variables. Examples of such forms of elaboration illustrate how the principle of parallelism can foster creativity in research design.

**The Utility of the Principle of Parallelism**

When crafting a study to estimate a treatment effect, researchers should consider all of the design options they can imagine that fit the demands of the research setting and then choose the most effective ones from among the identified possibilities. As I have tried to demonstrate, being mindful of the principle of parallelism can help researchers imagine a broader and richer range of methodological options from among which to choose, than they would otherwise, and thereby improve their research designs. The principle of parallelism might also be of use to methodological theorists who seek to explicate the general logic by which effects are estimated.

**References**


PRINCIPLE OF PARALLELISM


