A Typology of Strategies for Ruling Out Threats to Validity

Charles S. Reichardt

Don Campbell’s name will long be associated with the notion of threats to validity. In part, this is because he provided memorable and pedagogically appealing labels for, and categorizations of, threats to validity (Campbell, 1957; Campbell & Stanley, 1966; Cook & Campbell, 1979). Equally consequential have been Campbell’s demonstrations of the pernicious nature and omnipresence of threats to validity (e.g., Campbell & Boruch, 1975; Campbell & Erlebacher, 1970) and his explication of an epistemology whereby the effects of a treatment are assessed by ruling out plausible rival hypotheses (Campbell, 1974, 1988). The genius of Campbell’s (1969) contribution is also manifest in his imaginative illustrations of methods for ruling out threats to validity.

The present chapter, which I hope is faithful to the Campbellian tradition, provides a typology of strategies by which threats to validity can be ruled out. Researchers have long been ruling out threats to validity using a wide variety of methods. The typology I propose explicates the logic by which all these methods work. By understanding the logic behind their methods, researchers should be able to rule out threats to validity more effectively and, equally important, be better able to assess how effectively they have ruled them out.

AUTHOR’S NOTE: I am indebted to Len Bickman, Bob Boruch, Susan Grammer, Dave Kenny, Mel Mark, Jenny Novotny, Will Shadish, and Steve West for their helpful comments.
There are three general strategies for ruling out threats to validity: relabeling, substitution, and elaboration (Reichardt, 1983, 1988). In addition, the strategy of elaboration has a variety of subtypes, as depicted in Figure 5.1. Each of the three general strategies, plus the various subtypes of elaboration, are described below.

**RELABELING**

Threats to validity arise when a researcher estimates an effect of a treatment or intervention. For present purposes, estimating an effect will be conceptualized as an iterative process. This iterative process begins with a question specifying the effect to be estimated. The next step is to create a comparison by which to estimate the specified effect. The third step is to criticize the results of the first two steps. That is, the researcher considers how well the results of the comparison in fact estimate the effect specified by the question. Plausible criticisms are
Strategies for Ruling Out Threats to Validity

91
taken into account by cycling back to one of the two preceding steps. That is, the researcher recycles back either to change the comparison that was drawn or to change the question that was asked, which in turn might lead to a change in the comparison that was drawn. Then the researcher criticizes the results again, which can lead to a new round of changes in the question or the comparison, and so on. When repeated recyclings produce results for which apparent criticisms have been sufficiently addressed, the researcher exits with a conclusion.

Sometimes this iterative process is carried out as an intellectual exercise to help decide upon a study design. In such cases, the question and comparison that result during the final iteration are the only ones that are implemented. Other times, the iterative process is implemented in actions and not just thought; that is, the research project is carried out as a sequence of changing questions or comparisons.

A threat to validity is a criticism that can arise during the iterative process. In particular, a plausible threat to validity arises when it appears reasonably likely that the cause of an effect has been mislabeled.1 For example, consider a researcher who attempts to estimate the effect of a treatment by comparing the performance of a group of individuals who choose to receive the treatment to the performance of another group of individuals who choose not to receive the treatment. Furthermore, suppose the researcher attributes the mean outcome differences between these two groups to the effect of the treatment when in fact some of the mean difference in outcomes is due to initial differences between these two self-selected groups of research participants. Then the researcher has mislabeled the cause of the effect because the observed difference is not due to the effect of the treatment alone, as the researcher has specified, but to the effect of the treatment plus the effect of initial selection differences between the groups.

The simplest (though not necessarily the most desirable) way to remove a threat to validity is to change the label that the researcher originally proposes. For example, in the preceding illustration, instead of concluding that the observed outcome difference is equal to the effect of the treatment alone, the researcher would conclude that the observed difference is equal to the joint effect of the treatment and the effect of initial selection differences between the groups. Changing the label for the cause of an effect in this fashion is called “relabeling.”

In the iterative process described above, relabeling means, in essence, that the researcher has cycled back and changed the question that is being answered. For example, in the preceding illustration, the researcher originally sought to answer the question “What is the effect of the treatment?” but by relabeling is now answering the question “What is the joint effect of the treatment and initial selection differences?”

In some cases, an answer to the revised question may be just as interesting as an answer to the original question, or the revised question may suggest novel
areas of investigation by leading a researcher to reconceptualize an artifact as an
effect worthy of study rather than as a nuisance to be controlled (McGuire,
1969). This seems to be what has happened, for example, with experimenter
expectancies and placebo effects. That is, experimenter expectancies and
placebo effects were originally viewed solely as artifacts to be avoided but
later became the focus of theoretical interest, and so were purposely induced
so they could be studied.

In other cases, even though a change in the question may not be completely
desirable, relabeling nevertheless may be a desirable strategy. For example, if
a researcher wants to know the effect of a social program in the absence of
financial incentives, but the program was in fact accompanied by financial
incentives, the results may still be useful when properly labeled, and it is usu-
ally far better to have the result labeled properly than improperly.

For ruling out many (if not most) threats to validity, however, the answer to
the revised question will be far less interesting than the answer to the original
question. For example, it is far less interesting to know the size of the joint
effects of a treatment and of selection differences than to know the size of the
effect of the treatment alone. As a result, relabeling often will be a relatively
undesirable strategy for ruling out a threat to validity. When this is the case,
removing a threat to validity by altering the label will be a strategy of last
resort, to be used when neither of the other two strategies can be employed
effectively. Nonetheless, the relabeling strategy is a good option to keep in mind
because it can always be applied. If a threat to validity cannot be taken into
account in any other way, the threat can always be removed by relabeling.

SUBSTITUTION

The second strategy for ruling out a threat to validity is to replace the compari-
son from which the estimate of the treatment effect is derived. That is, a threat
to validity can be removed if the comparison that is subject to the threat is
replaced by a comparison that is not subject to the threat. Ruling out a threat
to validity in this manner is called “substitution,” because one comparison is
substituted for another.

For example, consider an investigator who plans to use a nonequivalent group
design (wherein individuals are assigned to treatment groups nonrandomly)
because of its ease of implementation but in which the effects of the threat to
validity of selection differences are likely to be large. After some reflection, the
researcher realizes that although an interrupted time series design would be
substantially more difficult to implement, the threat of initial selection differ-
ences would be avoided and the threats to validity, such as those due to history or
instrumentation that would be present, are likely to be small. In this case, the
researcher could rule out the threat to validity of initial selection differences
(and reduce the size of threats to validity overall) by using the interrupted time series design in place of the nonequivalent group design. Another example would be using a randomized experiment in place of a nonequivalent group design.

Campbell (1969) called threats to validity such as placebo effects, evaluation apprehension, and hypothesis guessing “instrumental incidentals” because they accompany but are theoretically irrelevant to the treatment under study. Campbell (1969) also distinguished between two approaches for taking account of the effects of instrumental incidentals, both of which are instances of substitution. His first approach was to replace the control group with an “expanded-content control group.” Using a placebo control condition in place of a no-treatment control condition is an example. Campbell’s second approach was to replace the experimental treatment with an “altered experimental treatment.” Adding deception to the experimental manipulation so as to avoid evaluation apprehension would be an example. Not all instances of substitution fall into one of these two categories; for example, switching from a nonequivalent group design to an interrupted time series design does not neatly fit into one of these categories.

Under some circumstances, it may not be possible to create a comparison via substitution wherein the threat to validity is completely removed. Nonetheless, it may still be possible in many cases to take account of the threat to validity with substitution by using a pair of comparisons. The technique would work in the following manner. One of the comparisons would be constructed so that the threat to validity was present and had a positive effect (i.e., so that the observed difference, assuming there were no other threats operating, would be greater than the true treatment effect). The other comparison would be constructed so that the threat to validity was present and had a negative effect (i.e., so that the observed difference, assuming there were no other threats operating, would be less than the true treatment effect). Then the size of the treatment effect would be bracketed within the two estimates (assuming no other threats were operating). Such a range of scores has been called a “plausibility bracket” (Reichardt & Gollob, 1987).

For example, suppose a researcher is not confident that a comparison can be created wherein the effects of hypothesis guessing are completely removed but is confident that two comparisons of the following nature can be constructed. In one comparison, the research participants would be led to believe that the hypothesis under study predicts that the experimental group would outperform the control group. In the other comparison, the research participants would be led to believe that the hypothesis under study predicts that the experimental group would underperform the control group. If the results of both comparisons are in the same direction (both showing, say, higher performance for the experimental group than for the control group), then the researcher can conclude that
the effects of hypothesis guessing have been taken into account (if there are no other threats to validity). That is, in the absence of other threats to validity, the effect of the treatment in this case is positive and is at least as large as the smaller estimate. Such a procedure will be called “bracketing by substitution.”

ELABORATION

The third strategy for ruling out a threat to validity is “elaboration” (Reichardt, 1995; Reichardt & Mark, 1998). Elaboration can perhaps best be described by contrasting it with substitution. Substitution takes account of a threat to validity by replacing the comparison used to estimate the effect. Instead of replacing the comparison as in substitution, elaboration takes account of a threat to validity by retaining the original comparison and adding another comparison. The additional comparison allows the researcher to disentangle the relative contributions of the treatment and of the threat to validity as explanations for the results in the original comparison.

For example, Aronson and Mills (1959; Aronson & Carlsmith, 1968) were interested in the effects that a stressful initiation into a group had on subsequent liking for the group. Half the sample of female college students were initiated into the group by reading a list of obscene words, which was thought to be stressful, whereas the other half entered the group with a much less stressful initiation. The results revealed that the more stressful the initiation, the higher was a subsequent rating of the attractiveness of the group. It is possible, however, that reading the obscene words was sexually arousing and that it was this arousal, rather than the stress, that caused the increase in the judged attractiveness of the group. To choose between these two explanations, Aronson and Carlsmith suggested assessing changes in the research participants’ emotional state. If the initiation of reading obscene words was stressful as intended, the emotional state of the participants would be expected to shift negatively, whereas if the initiation was sexually arousing, the emotional state would be expected to shift positively. In this way, adding an assessment of the direction of the shift in emotional state would allow the researcher to choose between the two alternative explanations.

As the preceding example illustrates, the comparison that is added in elaboration (i.e., the second comparison) is generated by thinking through the implications of the treatment and of the threat to validity so as to discover instances where these two sets of implications differ, then drawing a comparison to see the extent to which the differing predictions hold true. The label of “elaboration” comes from a remark that Cochran (1965, p. 252) attributed to Fisher:

When asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: “Make your theories elaborate.” The reply puzzled me at first, since by Occam’s razor the advice usu-
ally given is to make theories as simple as is consistent with the known data. What Sir Ronald meant, as the subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.

Elaboration can take a variety of forms. Six variations are described next. The first five variations are instances of "competitive" elaboration, and the last is "noncompetitive" elaboration (see Figure 5.1). The distinction between competitive and noncompetitive elaboration is defined after the first five variations are described.

**Elaboration 1: Show That the Size of the Effect of the Threat Is Zero**

Perhaps the simplest form of elaboration is one in which the second comparison reveals that the effect of the threat to validity is zero. For example, suppose a researcher wishes to assess the effects of meditation on a subsequent performance measure. Toward this end, the performance of two groups of research participants is compared where one group meditates and the other group listens to quiet music. A critic might speculate that the experimental group was not meditating but sleeping, and as a result the treatment was mislabeled; the effect assessed was not of meditation but of sleep. To take account of this alternative explanation, the researcher could add an assessment of the participants’ brain waves using an electroencephalogram (EEG). If the EEG recordings of all the participants in the experimental group reveal alpha waves, the participants were all awake and the alternative explanation is ruled out.

Of course, elaboration can be a double-edged sword. If it is the implications of the treatment, rather than of the threat to validity, that do not hold true in the additional comparison, the researcher has ruled out (or at least shed doubt on) the treatment rather than the threat as an explanation for the original results. For example, if the EEG recordings all reveal delta waves, the subjects were asleep and it is the alternative explanation that is supported while the meditation label is refuted. On the other hand, if the treatment and the threat to validity both remain tenable following the second comparison (e.g., some participants show alpha waves and some show delta waves), some other strategy must be used to take account of the threat to validity, perhaps another elaboration strategy.

**Explicit Versus Implicit Elaboration**

The strategy of showing that the size of the effect of a threat to validity is zero can be implemented either explicitly (as in the preceding example) or implicitly.
In the explicit form of the strategy, a threat to validity is articulated a priori. Then a comparison is added in which the given threat to validity is expected to produce a specific outcome, if it is indeed operating. If that expected outcome is not observed (assuming there are no other threats to validity operating), the given threat to validity is ruled out.

In the implicit form of the strategy, no specific threat to validity is articulated a priori. Instead, the researcher simply adds a comparison that assesses an additional implication of the treatment. Whatever threats to validity are incompatible with the results of the additional comparison are thereby ruled out.

Obviously, explicit elaboration is superior to implicit elaboration for ruling out a threat to validity that has been specified a priori. Because implicit elaboration does not focus on any given threat to validity, the implications of the treatment that are assessed may not directly bear on a threat to validity that is of particular interest and so may be ineffective in addressing that threat. In this regard, Platt’s (1964) well-known advocacy of “strong inference” can be conceptualized as an explanation of why explicit elaboration is often superior to implicit elaboration. Nonetheless, there will always be threats to validity that researchers have not thought of at any given point in time. Explicit elaboration cannot be applied to these as-yet-unspecified threats to validity, so implicit elaboration is the only elaboration approach that can be used.

The double-edged possibility of ruling out the treatment rather than threats to validity is especially apparent with implicit elaboration because the focus is only on implications of the treatment rather than on implications of a specified threat. For example, consider the report by Obmascik (1989) that elevated cancer rates in a city were initially blamed on pollution from a neighboring nuclear arsenal. Cancer rates in other nearby communities that received even greater pollution from the arsenal (i.e., an additional comparison in an implicit elaboration), however, were found not to be unusually high. This weakens the credibility of the original assertion about the effect of pollution more than it affirms any threat to validity. Similarly, if the cause of an effect is described as a treatment consisting of 20 hours of services but additional observation reveals that only 10 hours of services were delivered, one has demonstrated that the original label for the treatment was invalid more than one has shed doubt on any threat to validity. It is partly to avoid such mislabelings that the quantity and types of services received by research participants are assessed in applied research and manipulation checks are performed in laboratory studies. Correctly labeling a treatment obviously requires knowing the types and amounts of the interventions that are introduced, and this knowledge often is obtained from additional comparisons and observations.

From Popper’s (1959, 1965; also see Meehl, 1967, 1978, 1997) perspective, the value of assessing additional implications of a treatment is that such assessments make the test of the treatment as an explanation for the results more strin-
gent and, therefore, help validate that explanation to the extent the implications hold true. As Cordray (1986) emphasizes, the value of assessing additional implications of a treatment is not only to rule out threats to validity but also to “rule in” the putative cause as a plausible explanation.

The Elaborateness of Elaborations

In general, when performing implicit elaboration (i.e., when assessing implications of the treatment without regard for specific threats to validity), the more complex and precise the pattern of implications of the treatment that are predicted and confirmed, the greater the number of threats to validity that are ruled out or at least rendered less plausible (Campbell, 1969, 1974, 1988; Cook & Campbell, 1979). Conversely, the more complex and precise are the implications of the treatment that are investigated, the more validating is the test of the treatment as an explanation, if the test is passed. In this vein, Scriven (1974, 1976) draws an analogy between assessing configurations of treatment implications and the methods a detective or coroner uses to determine the cause of a death. Scriven labeled this process the “modus operandi (MO)” method, while Abelson (1995, p. 184) renamed it the “method of signatures”:

The specification of a recognizable signature enhances the credibility of claims that particular underlying processes are operative, much as would a coroner’s report (Scriven, 1974) of seven different signs of a heart attack (and no signs of any other cause of death).

Elaboration 2: Estimate and Subtract the Size of the Effect of the Threat

To avoid confusion, I will distinguish between the two comparisons that are involved in elaboration by calling one the original (or first) comparison and the other the additional (or second) comparison. Also for simplicity of presentation, I will assume that only a single threat to validity is present.

As noted previously, when a threat to validity is present, it means that the estimate derived from the original comparison is not equal to the effect of the treatment as labeled. Instead, the estimate is equal to the treatment effect plus the effect of that threat to validity. That is,

\[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]

where “\( \text{Estimate}_1 \)” is the estimate of the treatment effect derived from the original comparison. For example, a researcher could attribute an observed outcome difference (e.g., \( \text{Estimate}_1 \)) to the effect of a treatment when some of
or all the difference in outcomes is due to initial selection differences between the treatment groups and not just to the effect of the treatment. In this case, the initial selection differences are the threat to validity.

The “estimate-and-subtract” strategy adds a comparison that is equal to the size of the effect of the threat to validity. That is, the estimate from the second comparison is

\[
\text{Estimate}_2 = \text{Effect of the Threat}. 
\]

To estimate the treatment effect free from the effects of the threat to validity, the second estimate is subtracted from the first estimate. In other words, the effect of the threat to validity is estimated and subtracted, hence the name of the strategy. The strategy of showing that the size of the effect of the threat to validity is zero that was described previously is a special case of the estimate-and-subtract strategy where the effect of the threat to validity is simply shown to be zero and, therefore, need not be subtracted. Table 5.1 summarizes these and other forms of competitive elaboration.

An interrupted time-series design where a control time series is used to take account of the threat to validity of history provides an example of the estimate-and-subtract strategy. In an interrupted time-series design, observations are collected repeatedly on an experimental group both before and after a treatment is introduced. For the sake of argument, suppose there is an abrupt change in the level of the observations in the experimental group immediately following the introduction of the treatment and that this abrupt change is used as the original estimate of the treatment effect. Further, assume that this abrupt change could be due either to the effect of the treatment or to the effect of other historical changes that took place at the same time as the treatment was introduced. For simplicity, suppose that this “history” effect is the only threat to validity that is present. Then, in the experimental group

\[
\text{Estimate}_1 = \text{Treatment Effect} + \text{History Effect}. 
\]

Now suppose observations are also collected on a control group at the same points in time. Further, suppose the control group is subject to the effect of the same historical events but that the control group is not given the treatment (and again that no other threats to validity are operating). Then an observed difference in the control group could be calculated in the same manner as it was calculated in the experimental group. This difference is \( \text{Estimate}_2 \) and

\[
\text{Estimate}_2 = \text{History Effect}. 
\]
Then if the estimate derived from the control group (Estimate 2) is subtracted from the estimate derived from the experimental group (Estimate 1), the result is equal to the treatment effect, free of the effects of the history threat to validity. In other words, any difference between the before-after change in the experimental group and the before-after change in the control group is due to the treatment and not the threat. As a special case, if there is no abrupt change in the control group observations, then the abrupt change in the experimental group is due solely to the effects of the treatment. More generally, if the difference in the control time series is smaller than the difference in the experimental time series, the researcher can conclude that the threat of history has been rendered implausible as an explanation for the entire difference in the experimental series.

\[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]
\[ \text{Estimate}_2 = \text{Effect of the Threat} = 0 \]

**TABLE 5.1** Five Forms of Competition Elaboration

1. *Show the size of the effect of the threat is zero*
   \[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]
   \[ \text{Estimate}_2 = \text{Effect of the Threat} = 0 \]

2. *Estimate and subtract the size of the effect of the threat*
   \[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]
   \[ \text{Estimate}_2 = \text{Effect of the Threat} \]

3. *Vary the size of the treatment effect*
   \[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]
   \[ \text{Estimate}_2 = A \times \text{Treatment Effect} + \text{Effect of the Threat} \]

4. *Vary the size of the effect of the threat*
   \[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]
   \[ \text{Estimate}_2 = \text{Treatment Effect} + B \times \text{Effect of the Threat} \]

5. *Vary both the size of the treatment effect and the size of the effect of the threat*
   \[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]
   \[ \text{Estimate}_2 = A \times \text{Treatment Effect} + B \times \text{Effect of the Threat} \]
   (where \( A \neq B \))
   or
   \[ \text{Estimate}_1 = A \times \text{Treatment Effect} + B \times \text{Effect of Threat} \]
   \[ \text{Estimate}_2 = C \times \text{Treatment Effect} + D \times \text{Effect of Threat} \]
   (where \( A/B \neq C/D \))
**Informal Elaboration**

Concluding, as done immediately above, that the effect of a threat to validity is not large enough to account for the entire observed difference in Estimate\(_1\), without formally reporting an estimate for the size of the treatment effect, will be called “informal” elaboration. Note that a researcher needs to be careful when taking account of two or more threats to validity using informal elaboration. The reason is that informal elaboration can work well for any single threat to validity but can fail when applied to two or more threats concurrently (Reichardt & Gollob, 1989). For example, suppose the size of Estimate\(_1\) is 10 and the size of the effect of the first threat to validity is 5. Then using informal elaboration, a researcher can say that the effect of the threat is not large enough to account for the entire observed effect. Now suppose a second threat to validity is also shown to have an effect of size 5. Again, using informal elaboration, the researcher can say that the effect of this threat is also not large enough to account entirely for Estimate\(_1\), but in so doing, the researcher has not shown that both threats together are too small to account for the observed difference in Estimate\(_1\). To take account of both threats, the researcher would have to quantitatively adjust the size of the treatment estimate for the effect of each threat to validity. That is, the researcher would have to repeatedly use a formal estimate-and-subtract strategy. In this example, the researcher would find that the two threats together can account completely for the observed difference.

**Bracketing**

For simplicity, the estimate-and-subtract strategy was described above as if Estimate\(_2\) is exactly equal to the size of the effect of the threat to validity; however, such precision will be relatively rare in practice. Usually the second comparison can only be used to set limits on the size of the effect of the threat to validity rather than to derive its size exactly. In other words, usually the second comparison will be able to estimate the size of the effect of the threat to validity only within a range. In turn, when the second estimate is subtracted from the first estimate, the size of the treatment effect will correspondingly only be estimated within a range. Under these circumstances, the size of the treatment effect is said to be bracketed within a range of estimates, and such a procedure will be called “bracketing by elaboration.”

There are two ways to use the estimate-and-subtract strategy to set limits on the size of the effect of a threat to validity (as opposed to estimating its size exactly), and thereby to bracket the size of the treatment effect. First, rather than create a second estimate that is equal to the exact size of the effect of the threat to validity, the second estimate is made either smaller or larger than that effect.
For example, a lower bound would be set on the size of the treatment effect by making the second estimate greater than the size of the effect of the threat to validity. In this case, the first and second estimates would be the following:

\[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]

\[ \text{Estimate}_2 > \text{Effect of the Threat}. \]

Taking the difference between \( \text{Estimate}_1 \) and \( \text{Estimate}_2 \) would produce a number that is less than the size of the treatment effect, assuming no other threats to validity are operating. For example, suppose the first estimate equals 10 and the second estimate, which is created to be greater than the size of the effect of the threat, is 3. Then the treatment effect is greater than 7, assuming no other threats to validity are present. An upper bound on the size of the treatment effect could be derived by making the second estimate less than the size of the effect of the threat to validity.

Campbell (1969) provided the following illustration of how a lower bound on the size of the treatment effect could be established by setting an upper bound on the size of the effect of a threat to validity. The example comes from a study of the effects of culture on the susceptibility to visual illusions (Segall, Campbell, & Herskovits, 1966). An alternative explanation for one of the outcome differences was that the visual illusions being studied were presented differently in the two cultures. The materials used to present the illusions were supposed to be held vertically and 4 feet from the research participants, but it was feared that this might have been easier to accomplish in suburban living rooms than in more rustic tribal locales. To control for this rival explanation, the researchers added variation in the manner of presentation in the suburban sample, with half the illusions held as specified and half held horizontally and only 1.5 feet from the participants, which “was thought to be more slovenly than any actual administration, but in the likely direction of deviation” (Campbell, 1969, p. 364). This variation in the presentation of the illusions produced differences in the results, but they were not nearly large enough to account for the differences between the cultures, thereby rendering this threat to validity implausible as an explanation for the entire result.

The second way to set limits on the size of the effect of a threat to validity using the estimate-and-subtract strategy, and thereby to bracket the size of the treatment effect, is best explained with an example. Consider an interrupted time-series design that uses a control time series to estimate and subtract the effects of history. Although the control time series is susceptible to the same history effects as the experimental time series, the researcher is not convinced that the size of the history effect in the control time series is identical to the size of the history effect in the experimental series. The researcher is confident,
however, that the history effect in the control series is at least half as large as the size of the history effect in the experimental series; therefore, if the size of the estimate in the control time series is doubled and subtracted from the estimate in the experimental time series, a lower bound for the size of the treatment effect is established, assuming no other threats are operating. In symbolic form, the two estimates are the following:

\[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]

\[ \text{Estimate}_2 > 0.5 \times \text{Effect of the Threat.} \]

Then \( \text{Estimate}_1 - 2 \times \text{Estimate}_2 \) is less than the treatment effect. More generally, if

\[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]

\[ \text{Estimate}_2 > A \times \text{Effect of the Threat} \]

then \( \text{Estimate}_1 - (1/A) \times \text{Estimate}_2 \) is a lower bound on the size of the treatment effect. An upper bound on the size of the treatment effect can be established in a parallel fashion.

**Confidence Intervals**

Consider a randomized experiment that consists of an experimental group and a comparison group. Further, suppose 15 individuals are assigned to each group; the mean outcomes in the experimental and comparison groups are 25 and 15, respectively; and the standard deviations of the outcome scores in the two groups are 3 and 2, respectively. Then the mean outcome difference between the two groups (10 = 25 – 15) is an estimate of the average effect of the treatment. This estimate is susceptible to a threat to validity of initial selection differences. The treatment groups were specified to be assigned at random, and random assignment removes any bias due to selection differences, but random assignment doesn’t remove all selection differences. In particular, there are still random selection differences between the treatment groups. Therefore, the mean outcome difference between the groups is

\[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of Random Selection Differences}. \]
The size of the effect of random selection differences can be estimated within a range using the standard error of the mean outcome difference. This standard error is derived from the standard deviations within the two groups (and from the other information given above) and is equal to 0.93. Therefore, a researcher could be 95% confident that the portion of the observed mean difference that is due to random selection differences is between +1.9 and –1.9 (where 1.9 equals 0.93 times the $t$ value of 2.048). In other words, a second comparison (Estimate$_2$), derived from the variability of the scores within the groups, has been used to estimate the size of the effect of random selection differences within the following range:

$$-1.9 \leq \text{Effect of Random Selection Differences} \leq 1.9.$$  

Subtracting this range of estimates of the size of the effect of random selection differences from Estimate$_1$ produces the following range of estimates of the treatment effect:

$$8.1 \leq \text{Treatment Effect} \leq 11.9.$$  

The result is that a researcher could be 95% confident that the treatment effect lies within this range, assuming there are no other threats to validity operating. The range of 8.1 to 11.9 is a 95% confidence interval. In this way, a confidence interval can be seen to be a special case of bracketing via the estimate-and-subtract strategy. That is, a confidence interval is a special case of a plausibility bracket where the bracket takes account of uncertainty resulting from random selection differences (Reichardt & Gollob, 1987).

**Elaboration 3: Vary the Size of the Treatment Effect**

The third form of elaboration involves holding the effect of the threat to validity constant across the two comparisons but varying the size of the treatment effect. In schematic form, this results in the following two estimates:

$$\text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat}$$

$$\text{Estimate}_2 = A \times \text{Treatment Effect} + \text{Effect of the Threat}.$$  

A difference between the two estimates is indicative of a treatment effect, assuming no other threats to validity are operating. For a more quantitative result, the treatment effect can be estimated by taking the difference between
Estimate₂ and Estimate₁ and dividing by (A – 1). If the value of A is known only within a range, instead of being known exactly, the treatment effect also can be estimated only within a range. This strategy for taking account of a threat to validity is called elaboration by varying the size of the treatment effect. Note that the estimate-and-subtract strategy is a special case of the strategy of varying the size of the treatment effect in which A = 0 in Estimate₂.

An example comes from a study that assessed the effect of raising the level of galanin above the level that is produced naturally in the body. Leibowitz and Kim (1992; Azar, 1994) injected galanin into the paraventricular nucleus (PVN) of a random sample of rats. Compared to no-injection controls, the injected rats gained weight; however, the weight gain could have been due instead to the trauma of the injection. This threat to validity was ruled out because another comparison was added wherein the experimental rats were injected not with additional galanin but with a substance that blocks the galanin production that occurs naturally. Compared to the no-injection controls, these rats lost weight. Because the effect of the threat to validity (which is the effect of receiving an injection) is the same in the two comparisons, but the estimates from the two comparisons differed dramatically (i.e., in one there was a weight gain while in the other a weight loss), the threat to validity cannot account for the results.

Similarly, a bias due either to yea- or nay-saying (i.e., a tendency either to agree or disagree regardless of the content of a question) can be controlled by wording half the responses on a questionnaire so that agreeing corresponds to a high level of the trait being assessed and wording the other half so that agreeing corresponds to a low level of the trait. In this way, the two separate halves of the questionnaire are constructed so that they would have opposite treatment effects but common threats to validity. Reversing the scoring on the second half of the items and averaging the results from the two halves produces a measurement of the trait that is free from these wording effects. Interrupted time-series quasi-experiments with removed or reversed treatments (Cook & Campbell, 1979) are examples of other design features that take account of threats to validity by varying the size of the treatment effect.

In planning the use of this form of elaboration, it is perhaps worth noting that the size of a treatment effect can vary across comparisons either because different groups of individuals receive different amounts of the treatment (as in the above example with galanin) or because different groups of individuals are differentially affected by the same amount of a treatment. An example of the latter comes from a study of the effects of mandated busing on “white flight” wherein a larger treatment effect would be expected for the wealthy than for the poor because the wealthy could more easily afford to shift their residence outside busing zones (Glass, 1988).
Elaboration 4: Vary the Size of the Effect of the Threat

Another form of elaboration arises when the effect of the treatment is constant across the two comparisons but the size of the effect of the threat to validity varies. In schematic form, the two estimates are the following:

\[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat} \]

\[ \text{Estimate}_2 = \text{Treatment Effect} + B \times \text{Effect of the Threat}. \]

The absence of a difference between the two estimates is indicative of a treatment effect, assuming no other threats to validity are operating. For a more quantitative result, the treatment effect can be estimated by (a) multiplying Estimate\textsubscript{1} by \(B\), (b) subtracting Estimate\textsubscript{2} from Estimate\textsubscript{1}, and (c) then dividing by \((B - 1)\). If the value of \(B\) is known only within a range, instead of being known exactly, the treatment effect is also estimated only within a range. This strategy for taking account of a threat to validity is called elaboration by varying the size of the effect of the threat to validity.

An example comes from Minton (1975) as described in Cook and Campbell (1979; also see Mark, 1986). Minton’s study assessed the effect of *Sesame Street* by comparing the performance of younger siblings who watched *Sesame Street* to the performance of older siblings when they were the same age as the younger siblings but who had not watched *Sesame Street* because it had not been on the air. Specifically, the data for the younger siblings were taken from years after *Sesame Street* began broadcasting (which was in 1969) whereas the data for the older siblings were taken from years before *Sesame Street* began. For example, the performance of a younger sibling when he or she was 4 years old, which was in 1970, say, would have been compared to the performance of an older sibling when he or she was 4 years old, which might have been in 1968. Because each younger sibling was a *Sesame Street* viewer by the time he or she was 4 years old while the older sibling could not have been a viewer at the same age, the difference between their performances would reflect the effect of the show. Despite of matching for age, differences between the sibling pairs still remain. In particular, the older siblings occupied a higher position in the birth order than the younger siblings. For example, only the older siblings could have been firstborns, while the younger siblings were more likely to be secondborns. Because birth order is related to intellectual performance even when age is held constant (Zajonc & Markus, 1975), differences in birth order are a threat to validity.
To rule out this threat to validity, Cook and Campbell (1979) suggested using two comparisons. The first comparison would contrast older siblings who were firstborns with younger siblings who were secondborns. The second comparison would contrast older siblings who were secondborns with younger siblings who were thirdborns. Because differences in intellectual performance between firstborns and secondborns tend to be substantially larger than differences in performance between secondborns and thirdborns, the size of the effect of birth order should vary substantially across these two comparisons while the size of the effect of *Sesame Street* presumably would be held relatively constant. If the estimates from the two comparisons were similar in size, therefore, it would suggest that most of the difference between the sibling pairs was due to *Sesame Street*.

Consider the case where a threat to validity might result because the responses on a questionnaire are influenced by social desirability. Such a threat to validity might be taken into account by dividing the items in the questionnaire into two groups and drawing a comparison separately in each. Specifically, suppose two sets of items could be created wherein social desirability should have a large effect in one set but a small effect in the other (e.g., based on the wording of the items), but the treatment would be expected to have much the same effect in both sets. Then a small difference between estimates from the two sets of items would suggest that the effects of social desirability are small and therefore that most of the observed effect is due to the treatment. Threats to validity such as those due to experimenter expectancies, volunteer effects, and evaluation apprehension could be ruled out in a similar manner, that is, by partitioning the data so that in some comparisons the effects of the threats were large while in others the effects were small. In creating such partitions, researchers could either capitalize on variation in the size of effect of the threat that arises after the fact or purposefully introduce variation a priori.

**Elaboration 5: Vary Both the Size of the Treatment Effect and the Size of the Effect of the Threat**

In a more general form of elaboration, the effects of both the treatment and the threat to validity vary from one comparison to the other. In this case,

\[
\text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of the Threat}
\]

\[
\text{Estimate}_2 = A \times \text{Treatment Effect} + B \times \text{Effect of the Threat}
\]

where A and B can have any values as long as they are not equal; that is, the estimate from the second estimate is influenced by both the treatment and the threat to validity but in a different combination of amounts than in the first estimate.
If the values of A and B are known, then Estimate₁ and Estimate₂ can be combined so as to derive the size of the treatment effect, just as two algebraic equations can be solved simultaneously for the value of an unknown. Alternatively, if the sizes of A and B are known only within ranges, then the treatment effect can be estimated only within a range.

An even more general form of elaboration is the following:

\[ \text{Estimate}_1 = A \times \text{Treatment Effect} + B \times \text{Effect of Threat} \]
\[ \text{Estimate}_2 = C \times \text{Treatment Effect} + D \times \text{Effect of Threat} \]

where the ratio A/B is different from the ratio C/D. Many statistical procedures, including multiple regression, can be conceptualized as instances of this general form of elaboration.

So far it has been assumed that the effects of the treatment and of the threat to validity are additive rather than multiplicative. That is, in simplified form it has been assumed that

\[ \text{Estimate}_1 = \text{Treatment Effect} + \text{Effect of Threat} \]

where the equation contains an addition sign, rather than

\[ \text{Estimate}_1 = \text{Treatment Effect} \times \text{Effect of Threat} \]

where the equation contains a multiplication sign. Either type of combination, however, is possible. For example, the biasing effects of method variance have been modeled alternatively as additive or multiplicative in the analysis of data from a multitrait-multimethod matrix (Browne, 1989; Campbell & Fiske, 1959; Campbell & O’Connell, 1967, 1982; Reichardt & Coleman, 1995). If the effects of the treatment and of the threat to validity multiply rather than add, the logic of elaboration remains the same but the computational details are altered. Both space and the level of technical presentation precludes further discussion of the more complex forms of elaboration that are noted in the present section.

**Elaboration 6: Noncompetitive Elaboration**

The preceding five elaboration strategies are all instances of “competitive” elaboration (see Figure 5.1). The present section describes the alternative of noncompetitive elaboration and defines the difference between competitive and noncompetitive elaboration.

In competitive elaboration, two estimates are created in which the implications of the treatment differ from the implications of the threat to validity. Then,
in simplified form, the researcher assesses whether it is the implications of the treatment or of the threat to validity that hold true. In other words, a threat to validity is ruled out by putting the treatment and the threat to validity into competition as explanations for the results (Reichardt & Mark, 1998). In more general terms, competitive elaboration adds a comparison so as to separate or disentangle the effects of the treatment from the effects of the threat to validity as they are confounded in the original comparison.

In contrast, noncompetitive elaboration adds a comparison in which the treatment and the threat to validity have the same implications (i.e., predict the same outcome). If the implications of the threat to validity fail to hold true both the threat to validity and the treatment are thereby ruled out (assuming there are no other threats to validity operating). In this way, the treatment and the threat to validity are not placed in competition with each other; rather, the treatment and the threat to validity are codefendants at a joint trial of their guilt or innocence, and so are noncompeting.

Noncompetitive elaboration allows the researcher to investigate any implication of a threat to validity, whereas competitive elaboration restricts the researcher to investigating only implications of the threat to validity that are at odds with the implications of the treatment. If neither the treatment nor the threat to validity has an effect in the original comparison, noncompetitive elaboration can be superior to competitive elaboration. This is because noncompetitive elaboration allows the study of implications that might rule out the threat to validity but that would not be allowed with competitive elaboration. If either the treatment or the threat to validity, or both, has an effect in the original comparison, noncompetitive elaboration will be no better (and could well be more time-consuming) than competitive elaboration. Because usually either the treatment or the threat to validity (or both) has some effect, competitive elaboration tends to be more useful than noncompetitive elaboration in most cases. As a result, noncompetitive elaboration is relatively rare in practice. It is described here to provide a more complete overview of the logic of elaboration.

**SOURCES OF IMPLICATIONS FOR ELABORATION**

The elaboration strategy entails thinking through implications either of the treatment or of the threat to validity (or both) and adding observations to assess these implications. The implications to be assessed can come from many different sources, and there is much room for creativity in discovering or (as Scriven, 1974, emphasizes) creating useful implications. Below, just a few of the sources of implications that have been noted in the literature are mentioned.

Scriven (1974, 1976) and Mark (1990) emphasize assessing the presumed mediating or intervening processes of a putative cause. Toward this end, Scriven
proposes creating “signature arrangements,” “tracers,” and “tell-tales.” For example, in an evaluation of a teaching bureau on a college campus, Scriven (1976, pp. 109-110) suggests that the “procedures recommended by the bureau [be made] distinguishable from those coming from other sources” such as by having verbal counseling contain novel phrases, and then “monitoring the bloodstream of information through the university later, [so] we can detect the passage of ‘signed’ material and assess deterioration, implementation, and so forth.” At least partly for the same reasons, Aronson, Ellsworth, Carlsmith, and Gonzales (1990) suggest conducting manipulation checks and “internal analysis” of the presence of the treatment as labeled.

Both Abelson (1995) and Phillips and Bollen (1985) emphasize the importance of implications where the effect of a putative cause is expected to vanish under specified conditions:

The most compelling kind of supporting evidence for an initially implausible relationship involves selective variations in the focus of the analysis, such that the relationship should go away if the investigator’s hypothesized mechanism for the relationship is correct. (Abelson, 1995, p. 186)

An example comes from Phillips (1977), who found that well-publicized suicides (of any nature) were accompanied, the next week, by an increase in traffic fatalities. The proposed interpretation, that publicized suicides cause an increase in suicides on the highway, was strengthened by showing that there was no increase in traffic fatalities in cars with passengers (only in cars with lone drivers) and that there was no increase in such fatalities the week before publicized suicides, only the week after. Recognizing that null treatment effects are to be expected if one looks backward in time, as in looking for effects the week before publicized suicides, is a design-and-analysis feature that Campbell and Erlebacher also (1970) suggest.

Individual differences provide not only a possible but a necessary source of implications, according to Underwood (1975). That is, Underwood (1975, p. 130) argues that a theory’s implications about individual differences should be tested before the theory is allowed to “see the light of day.” For example, a theory that discrimination learning is based on frequency judgments implies that a test should be made to see if individuals who are better at making frequency judgments tend also to be better at learning discriminations. Implications such as that individuals who show the strongest “take” on manipulation checks should also exhibit the largest treatment effects are less theoretically sophisticated implications but are based on the same reasoning.

Aronson and Carlsmith (1968; Aronson et al., 1990) emphasize the value of both assessing multiple outcome measures and replicating results using novel implementations of the treatment manipulation, as does Campbell (1969).
For example, Aronson and Carlsmith (1968) note how Miller (1957) increases one’s confidence that hunger is the cause of an effect by assessing outcomes on a variety of measures including the “volume of food consumed, stomach contractions, rate of bar pressing to obtain food, and amount of quinine tolerated in food” (Aronson & Carlsmith, 1968, p. 16). Similarly, they argue that we can increase our confidence in the conclusion that a stressful initiation into a group increases liking for the group by replicating the study using a variety of different instantiations of stressful initiation tasks such as reading obscene words, doing push-ups, undergoing electric shocks, and the like.

Reichardt and Mark (1998) demonstrated how to find relevant implications by looking for differential effects of the treatment and the threat to validity across different causes, outcome variables, recipients, settings, and times. Reichardt (1992) noted that an additional comparison could be obtained either by collecting new data or by disaggregating existing data. For example, the Ross, Campbell, and Glass (1970) study of the effect of the British Breathalyser crackdown on drunken driving added a comparison by disaggregating available data according to the time during the week, looking separately at traffic fatalities when pubs in England were open and when they were closed. Similarly, the suggestion by Cook and Campbell (1979) with regard to Minton’s (1975) study of the effects of Sesame Street, as described previously, was to disaggregate the existing data into pairs of siblings who were first- and secondborns, second- and thirdborns, and so on.

**SUMMARY AND CONCLUSIONS**

A threat to validity can be taken into account in one of three ways: relabeling, substitution, or elaboration. Relabeling takes account of a threat to validity not by altering the size of the treatment effect estimate but by altering the label for the treatment effect. The advantage of this strategy is that its implementation presumes nothing more than that a threat to validity can be articulated. The drawback to relabeling is that it changes the question that is being asked, and often the new question will be far less interesting than the original. Unlike relabeling, the two other strategies for taking account of threats to validity (substitution and elaboration) provide means for altering the size of the treatment effect estimate rather than altering the label. The drawback to these two strategies is that they are more demanding to implement. In addition, substitution and elaboration often can be used to take account of a threat to validity only within a range of scores; that is, substitution and elaboration often will be able to provide only a range of estimates (rather than a single point estimate) of the treatment effect that is free of the effects of a given threat to validity. Producing a range of estimates for a treatment effect that is free from the effects of a specified threat to validity is called bracketing (Reichardt & Gollob, 1987).
Substitution takes account of a threat to validity by replacing the comparison that was originally used to estimate the treatment effect with a new comparison. In contrast, elaboration takes account of a threat to validity not by replacing the original comparison but by adding a comparison. A broad array of design and analysis features rule out threats to validity by means of elaboration. In general terms, the logic of elaboration is to create a pattern of outcomes so that the effects of the treatment and of the threat to validity can be disentangled. In implementing elaboration, there is often value in trying to assess implications of the treatment and of the threat to validity that are complex and precise. The purpose in so doing is to find patterns of outcroppings in the data that uniquely distinguish the causes of an effect in the same manner that a signature or fingerprint uniquely identifies individuals.

Recognition of the critical role played by patterns, especially complex patterns, is one reason why the elaboration strategy has often been called “pattern matching” (Campbell, 1966; Mark, 1990; Trochim, 1985, 1989). The advantage of the “elaboration” label is that it, more than “pattern matching,” connotes the critical role of the making, and not just the comparing, of patterns. The “elaboration” label is also more descriptive of the logic of ruling out threats to validity than the more common admonition to use “multiple methods.” Multiple methods can be used for many purposes besides elaboration, and elaboration can often get by with just a single method used repeatedly (e.g., in different settings or with different recipients). In addition, “multiple” provides the wrong emphasis when the focus is on bracketing an effect size. Bracketing is better accomplished using just two estimates wherein the direction of bias is known to differ, which is the basis of elaboration, than by using multiple estimates wherein the biases differ in source (but are not necessarily known to differ in direction), which is typically the espoused goal when using “multiple methods.”

Confidence intervals are a form of bracketing via elaboration. Statistical significance tests also rely on the logic of bracketing; however, confidence intervals and statistical significance tests only bracket threats to validity resulting from random variation. Brackets are also needed for threats to validity that are not random (Reichardt & Gollob, 1986, 1987, 1989). Although confidence intervals and statistical significance tests are widely used, brackets to take account of nonrandom threats to validity are not. That is, rather than taking account of a nonrandom threat to validity within a range, researchers often use a single, point estimate, even in the face of obvious and substantial uncertainty about the correctness of the estimate. For example, we often do not know how to adjust properly for the effects of initial selection differences between non-equivalent groups (Cochran & Rubin, 1973; Lord, 1967; Reichardt, 1979; Rosenbaum, 1995), but researchers often use a single analytic approach as if it were exactly correct. Even when researchers use more than one approach so
as to produce a range of estimates, they often have little sense of whether the range actually brackets the likely size of the treatment effect. Consequently, it is likely that the degree of uncertainty that is present in the results is often underrepresented and that researchers therefore are misleading their readers, if not themselves. Although brackets that are both adequate and useful can be extremely difficult to construct in practice, we should, at the very least, be teaching their use both as the ideal toward which we should be striving and as the appropriate standard against which we should assess the adequacy of our methods. Perhaps continuing to develop a logic for taking account of threats to validity will help us achieve these goals.

NOTES

1. The present chapter will be concerned only with mislabelings of the cause. According to the definitions in Cook and Campbell (1979), mislabelings of the cause entail either threats to internal validity or threats to the construct validity of the cause. Nonetheless, the strategies described herein are applicable to all threats to validity, if care is taken to distinguish a threat to validity from other forms of criticism (Reichardt, 1983, 1988).

2. In the same vein, the expression “taking account of threats to validity” is often more appropriate than the expression “ruling out threats to validity.” “Taking account of” seems compatible with the notion of bracketing and its accompanying degree of uncertainty, whereas “ruling out” suggests that the presence of a threat to validity has been completely exorcised and, therefore, that the appropriate outcome is a point estimate rather than a range. “Ruling out” also seems, more than “taking into account,” to suggest that the only appropriate goal in addressing a threat to validity is to show that its effect is zero rather than to adjust for the size of its effect, whatever it may be, as can be accomplished with elaboration. Nonetheless, “ruling out” is widely used and accepted, and I have succumbed to its use at many times in the present chapter for ease of expression.

REFERENCES


