PRIMER

Establishing the internal and external validity of experimental studies

MARION K. SLACK AND JOLAINE R. DRAUGALIS

Abstract: The information needed to determine the internal and external validity of an experimental study is discussed.

Internal validity is the degree to which a study establishes the cause-and-effect relationship between the treatment and the observed outcome. Establishing the internal validity of a study is based on a logical process. For a research report, the logical framework is provided by the report's structure. The methods section describes what procedures were followed to minimize threats to internal validity, the results section reports the relevant data, and the discussion section assesses the influence of bias. Eight threats to internal validity have been defined: history, maturation, testing, instrumentation, regression, selection, experimental mortality, and an interaction of threats. A cognitive map may be used to guide investigators when addressing validity in a research report. The map is based on the premise that information in the report evolves from one section to the next to provide a complete logical description of each internal-validity problem. The map addresses experimental mortality, randomization, blinding, placebo effects, and adherence to the study protocol. Threats to internal validity may be a source of extraneous variance when the findings are not significant. External validity is addressed by delineating inclusion and exclusion criteria, describing subjects in terms of relevant variables, and assessing generalizability.

By using a cognitive map, investigators reporting an experimental study can systematically address internal and external validity so that the effects of the treatment are accurately portrayed and generalization of the findings is appropriate.

Index terms: Clinical studies; Control, quality; Methodology; Researach

Am J Health-Syst Pharm. 2001; 58:2173-84

The effects of investigational treatments are established by statistically testing the findings to determine if any differences are likely to be due to chance alone and by examining the study's design and execution to rule out alternative causes of the observed effects. The process of ruling out alternative causes is referred to as assessing or establishing internal validity. Internal validity is the degree to which a study establishes the cause-and-effect relationship between the treatment and the observed outcome; conversely, it refers to the degree to which the absence of a relationship implies the absence of cause.[1,2] Internal validity is the sine qua non of research; without it, a study is meaningless.[1] In a study that lacks internal validity, the results are probably attributable to a cause other than the treatment. Consequently, one could not expect to observe similar effects if the study was duplicated, nor could the results be generalized to similar populations.

Internal validity, as defined by Campbell and Stanley,[1] is a logical rather than statistical issue. Statistical tests establish the likelihood that the study results are due to chance variation rather than to the treatment or some other cause. When the result is not likely attributable to chance (i.e., the value of p is 0.05 or less), then the design and execution of the study are assessed to judge whether the effect resulted from the treatment or from another factor. Even if the statistical test does not indicate significance (i.e., p is greater than 0.05), the design and execution of the study can still be assessed to determine if extra-

MARION K. SLACK, PH.D., is Research Scientist/Teaching Associate and JOLAINE R. DRAUGALIS, PH.D., is Professor and Assistant Dean, College of Pharmacy, The University of Arizona, Tucson.

Address correspondence to Dr. Slack at the College of Pharmacy, The University of Arizona, P.O. Box 210207, Tucson, AZ 85721-0207 (slack@pharmacy.arizona.edu).

Presented at the ASHP Midyear Clinical Meeting, Orlando, FL, December 8, 1999.

This is **article 204-000-01-012-H04** in the **ASHP Continuing Education System;** it qualifies for 1.0 hour of continuing-education credit. See page 2182 or http://ce.ashp.org for the learning objectives, test questions, and answer sheet.

The Primer *section covers basic information in various fields of knowledge of interest to pharmacists who practice in health systems. Within the scope of the section are reviews of fundamental concepts in, for example, pharmacy, pharmaceutics, pharmacology, physiology, therapeutics, and health care technology. Also covered are topics somewhat out of the mainstream of pharmacy (e.g., advances in nondrug health care technology) but nevertheless of interest to practitioners.*

neous factors obscured the treatment's impact. To effectively establish that a treatment produced the outcome, the investigator must show that extraneous factors were unlikely to have influenced the results.

The types of extraneous factors that can influence the outcome of a study depend on the research design.[1,3] That is, the extraneous factors that can affect the outcome of a true experimental study are different from those that can influence a study involving a pretest–posttest design and a single group of subjects. A true experimental design is one that has at least two independent, parallel groups; randomly assigns subjects to the groups; and assesses treatments prospectively.

Studies evaluating experimental research designs have shown that poor execution of specific study procedures can bias results. Schulz et al.[4] examined the association between treatment effects and procedures such as allocation concealment, sequence generation, withdrawals ("dropouts"), and blinding. They reported that the effects of treatment were 30% greater in studies with inadequate allocation concealment than in studies with adequate concealment. Similar results were observed in studies that lacked appropriate blinding. Less bias was attributed to sequence-generation procedures or to dropouts.

This article delineates the methodological issues associated with experimental research designs, shows how they differ from those associated with other designs, and provides a cognitive map for investigators to use to ensure that they address pertinent methodological issues when reporting the results of an experimental study and for readers to use when determining if a report adequately addresses internal validity. In addition, we describe the implications of methodology for study outcomes when the results are not significantly different among study groups and discuss external validity, or generalizability.

Theoretical framework

The following discussion is based on the framework of Campbell and Stanley,[1] in that specific threats to establishing a cause-and-effect relationship (i.e., internal validity) are associated with the particular research design and with how the study procedures are executed. Therefore, the investigator needs to know which threats to internal validity are associated with which research designs and the sources of bias associated with particular aspects of study execution.

We also follow Campbell and Stanley's contention that establishing the internal validity of a study or assessing bias is based on a logical process. Hence, the information needed to assess internal validity must be presented so that the reader has the critical information available in a logical sequence. For a research report, the logical framework is provided by the report's structure. The methods section describes how the study was designed and what procedures were followed to reduce or eliminate specific threats to internal validity. The results section reports the data relevant to establishing internal validity, and the discussion section provides the investigators' assessment of the influence of bias. For a specific threat to internal validity or source of bias, a logical thread of information should be readily identifiable that progresses from the methods section to the results section to the discussion section.

This approach differs from a scale or checklist method, in which the reader scores a research report for a number of methodological factors. Scales and checklists are usually concerned with overall study quality and have been criticized for not having a theoretical foundation.[5] They lack logical order and list methodological factors that are not specific to the research design. For example, carryover effects, a consideration in a crossover design, may be included on the list, even though the study has a randomized design with parallel groups. Also, checklists typically assess multiple types of validity; they contain items related to statistical conclusion validity and external validity, as well as items related to internal validity.[6]

Threats to internal validity

Internal validity is concerned with the rigor (and thus the degree of control) of the study design. The degree of control exerted over potential extraneous variables determines the level of internal validity. Controlling for potentially confounding variables minimizes the potential for an alternative explanation for treatment effects and provides more confidence that effects are due to the independent variable. Eight threats to internal validity have been defined: history, maturation, testing, instrumentation, regression, selection, experimental mortality, and an interaction of threats.[1,2]

History. History becomes a threat when other factors external to the subjects (in addition to the treatment variable) occur by virtue of the passage of time. For example, the reported effect of a year-long, institution-specific program to improve medical resident prescribing and order-writing practices may have been confounded by a self-directed continuing-education series on medication errors provided to resi-

dents by a pharmaceutical firm's medical education liaison.

Maturation. The maturation threat can operate when biological or psychological changes occur within subjects and these changes may account in part or in total for effects discerned in the study. For example, a reported decrease in emergency-room visits in a long-term study of pediatric patients with asthma may be due to outgrowing childhood asthma rather than to any treatment regimen imposed. Both history and maturation are more of a concern in longitudinal studies.

Testing. The testing threat may occur when changes in test scores occur not because of the intervention but rather because of repeated testing. This is of particular concern when researchers administer identical pretests and posttests. For example, a reported improvement in medical resident prescribing behaviors and order-writing practices in the study previously described may have been due to repeated administration of the same short quiz. That is, the residents simply learned to provide the right answers rather than truly achieving improved prescribing habits.

Instrumentation. When study results are due to changes in instrument calibration or observer changes rather than to a true treatment effect, the instrumentation threat is in operation. For example, in a communications course, evaluator 1 observes pharmacy students counsel a patient at week 3 of the semester, and evaluator 2 observes the students at the conclusion of the course. If the evaluators are dissimilar enough in their approach, perhaps because of lack of training, this difference may contribute to measurement error in trying to determine how much learning occurred over the semester.

Regression. The regression threat can occur when subjects have been selected on the basis of extreme scores, because extreme (low and high) scores in a distribution tend to move closer to the mean (i.e., regress) in repeated testing. For example, if a group of subjects was recruited on the basis of extremely high stress scores and an educational intervention was conducted, any postintervention improvement noted could be due partly, if not entirely, to regression rather than to the coping techniques presented in the educational program.

Differential selection. The selection threat is of utmost concern when subjects cannot be randomly assigned to treatment groups, particularly if groups are unequal in relevant variables before treatment intervention. For example, one obstetrics and gynecology clinic's patients receive a pharmacy-based educational intervention and another clinic's patients receive a mailed pamphlet; both methods are designed to encourage calcium supplementation. When the outcome is measured at the end of the study, it may be confounded by the fact that the groups were not equal with respect to relevant variables (e.g., age, currently provided educational materials, hysterectomy status, menopausal status) before the educational program was implemented.

Experimental mortality. Experimental mortality is also known as attrition, withdrawals, or dropouts and is problematic when there is a differential loss of subjects from comparison groups subsequent to randomization, resulting in unequal groups at the study's end. One example is a study designed to compare the effects of an intranasal corticosteroid spray with placebo in alleviating symptoms of allergic rhinitis. If subjects with the most severe symptoms preferentially dropped out of the active treatment group, the treatment may appear more effective than it really is.

Selection interactions. The final threat to internal validity is an interaction of the selection threat with any of the other threats. The selection interaction most commonly confronted involves maturation. The selection–maturation interaction concerns the differential assignment of subjects to groups in a way that relates to the subjects' maturation. For example, two groups of diabetic patients may have similar disease indicators at the start of a study, yet a treatment effect could result if a larger percentage of patients in whom an effect of maturation (e.g., progressive worsening of disease) is more prevalent are assigned to one group.[7]

The research design chosen (e.g., experimental, quasi-experimental, one-group pretest–posttest) and operational procedures used (e.g., randomization techniques, adherence standards) determine the level of confidence in the internal validity. Knowledge of the potential threats and the ability to discern to what degree they may be operating in a study enable one to better analyze the results.

Random assignment to parallel groups, the hallmark of an experimental study, effectively controls all threats to internal validity except experimental mortality. Differential selection is controlled because random assignment creates groups that are equivalent with respect to known and unknown variables so that differences in outcomes cannot be caused by differences among groups. Other threats, for example maturation, are ruled out by the presence of one or more parallel groups. Because maturation should occur equally in all the groups, any difference in response should be due to the treatment. No other research design can control for so many threats at once. This is why experimental studies are considered the standard of research design.

Cognitive map for establishing the effects of treatment

Cognitive maps are plans or procedures for completing a task or accomplishing a goal.[8] A cognitive map provides a skeleton for directing the analytical process and guiding the

logic of the writing; it also provides rules for organizing the final product and facilitates systematic examination of issues.[9] Such a tool is believed to be important to analytical thinking. We developed a cognitive map to guide investigators when addressing validity issues in a research report.

The cognitive map shown in Table 1 is based on the premise that each section of a research report provides specific information related to establishing the effects of a treatment and that the information evolves from one section to the next to provide a complete logical description of each internal-validity problem. In the table, the components proceed from left to right; information evolves from a description in the methods section of study procedures intended to prevent or limit design or methodological problems to a report in the results section of findings relevant to establishing internal validity to an assessment in the discussion section of the impact of any internal-validity problems on study outcomes. Throughout this part of the discussion, we assume that the findings were statistically significant, that is, that differences among groups are probably not due to chance variation. The introduction and conclusion sections do not provide direct information on internal validity and are not included.

From a practical perspective, the central issue in demonstrating internal validity and establishing the effects of a treatment is ensuring that the comparison groups (the treatment and control groups) are equivalent in all variables except the independent (treatment) variable. In other words, the groups are similar demographically and do not differ in severity and type of disease, prognosis, or comorbidities and in how they were handled during the study, except for the experimental treatment.

Experimental mortality. The first internal-validity factor listed in Table 1 is experimental mortality. To reiterate, experimental mortality involves any subject who has been enrolled in a study and randomly assigned to a group but not included in the analysis for any reason.[10] Participants may be excluded from the analysis for a number of reasons, including ineligibility (subjects admitted to study because of clerical or diagnostic errors), nonadherence to the study protocol (by either subjects or researchers), poor or missing data, and competing events.

Because the value of random assignment is lost if subjects are dropped from the analysis (the groups can no longer be considered equivalent in terms of known and unknown factors), the preferred procedure for preventing bias is an intention-to-treat analysis, in which

Table 1.
Cognitive Map for Establishing Internal Validity of Experimental Studies

Internal-Validity Factor	Information in Section of Research Report[a]		
	Methods Section	Results Section	Discussion Section[b]
Related to Study Design			
Experimental mortality	Description of data analysis for study dropouts, or use of intention-to-treat analysis or appropriate statistical analysis	Demographics and clinical outcomes tables: statistical tests used to compare baseline characteristics and dependent variable between groups consistent with intention-to-treat analysis, or analysis with and without data from dropouts	Reasons for withdrawal reported. If intention-to-treat analysis not used, discusses impact of dropouts on data interpretation and dependent variable
Related to Study Procedures			
Randomization	Description of randomization method, baseline data collected, and statistical analysis of baseline data	Demographics table: statistically compares study groups in terms of relevant demographic data	Differences between groups and their impact on results discussed
Blinding	Description of blinding procedures; if no blinding, discussion of methods used to prevent bias	Effectiveness of blinding reported; if no blinding, data showing treatment equivalence (except with respect to independent variable) reported	Issues related to blinding and their impact on results discussed
Placebo	Description of matching placebo, discussion of effects related to placebo	Assessment of subjects' and providers' knowledge of treatment	Issues related to placebo and their impact on results discussed
Adherence to protocol	Description of methods used to assess adherence and of adherence standards	Protocol adherence for all treatment groups reported	Compliance issues and their impact on findings discussed

[a] Introduction and conclusion sections are not included, since they do not provide direct information on establishing the effect of treatment.
[b] In general, threats to internal validity are not addressed in the discussion section if the methods and results sections establish that the threat is unlikely to play a role in the study.

all subjects randomized are included in the analysis.[3,11] Although the exact reasons for withdrawal from the study do not affect an intention-to-treat analysis, they may be informative for future studies or when using the treatment in practice.

In a simple intention-to-treat analysis, all subjects are retained in the denominator if the dependent variable is a proportion (e.g., the proportion of patients who improved) and the last obtained measurement is used for a continuous variable (e.g., blood pressure). Investigators should state whether they used intention-to-treat analysis in the methods section.

If an intention-to-treat analysis is not used, then the analysis that was used must be described and the investigators must verify that no bias was present as a result of withdrawals. If there was bias, the investigators must discuss its impact on the estimate of treatment effect. In general, establishing that withdrawals did not bias the findings is much more onerous than using an intention-to-treat analysis. The investigator must show that the analysis was not biased and that subjects did not withdraw differentially from the study groups.[10] Although information on the relative number of dropouts from each group and the reasons for withdrawal may provide insight into the causes of experimental mortality, such information does not establish equivalence for unknown factors, nor does it rule out the possibility that dropouts are related to treatment. Thus, alternative methods of analysis are always less desirable, and the results more tentative, than if intention-to-treat had been used.

In the results section, investigators establish that an intention-to-treat analysis was indeed used by showing that the number of subjects randomized to study groups was the same as the number of subjects for whom baseline data and outcomes data were reported. The baseline demographic data and the outcomes data are typically presented in two separate tables. In such cases, the total number of subjects in each table should match the total number of patients randomized. For example, if the authors state that 309 women were enrolled in the study, the total number of patients in the demographics table must equal that in the outcomes table (n = 309).

The next four internal-validity factors listed in Table 1 are related to the implementation of study procedures. Procedures such as random assignment, double-blinding, using a placebo, and using protocols should prevent bias from influencing measures of the dependent (outcome) variable. However, they must be implemented correctly; carelessly executed procedures are common sources of bias. We now describe what information is needed to determine if the study procedures were implemented in a manner that did not introduce bias.

Randomization. Randomization is the first study procedure outlined in Table 1. Note that randomization, or random assignment, is a different process with a different objective than random selection. Random assignment uses a random process, such as a coin toss, a table of random numbers, or computer-generated random numbers to determine the type of treatment (e.g., drug or placebo) that each study participant receives. Random selection uses a random process to identify study participants from the population. Because random assignment is related to internal validity and random selection to external validity, the two procedures should not be confused.

Randomization is the best method available to produce study groups that are equivalent with respect to known and unknown variables.[3,10] However, the randomization procedure must be executed in a manner that does not introduce bias into the study. Recommendations include identifying the method of randomization used, the method used to conceal the assignment schedule until recruitment is complete, who generated and who executed the allocation scheme, and relevant baseline data showing that the study groups are equivalent in terms of known variables.[12]

Concealing the allocation sequence from providers who enter subjects into a study appears particularly important. That is, the provider should not know which treatment the next subject would receive if admitted into the study. Concealment prevents bias from entering into the process of determining subject eligibility and assigning treatment. Studies without concealment of the allocation sequence find effects 30% larger than studies with concealment.[4,13]

Statistical tests are used to compare the baseline variables of all treatment groups. This establishes that the random-assignment procedure indeed resulted in groups that were similar for measured variables and that bias resulting from the randomization process was unlikely. Authors may report p values when comparing baseline variables among study groups; however, a p value indicates if the randomization was fair, not whether the groups were equivalent. Therefore, the prognostic strength of the variables and the magnitude of the difference also need to be considered.[14] If the groups are not equivalent for all variables, the differences should be addressed in the discussion section and the impact of the differences on the reported outcomes judged. In one study, analysis of baseline characteristics revealed differences between groups in exposure to smoke, fat intake, and alcohol consumption.[15] The investigators then used multivariate logistic regression to assess the impact of the unequal groups on the results of the study. The regression analysis supported their contention that differences between the groups were not responsible for the findings.

Blinding. If a study is blinded, the procedures used to blind patients and providers to treatment assignments should be described in the methods section, any data on the effectiveness of the blinding should be reported in the results section, and any relevant issues should be addressed in the discussion section. A study that evaluated physicians' interpretation of blinding found substantial variability between readers' interpretations and textbook definitions of the terms "single blind," "double blind," and "triple blind."[16] Therefore, it was suggested that authors specifically state the blinding status of everyone involved in a study. Providing data on the effectiveness of blinding is particularly important if characteristics of the treatment allow subjects to identify whether they are receiving the drug or placebo. For example, in a study comparing zinc and placebo lozenges, the investigators asked subjects to guess their study assignment.[17] They reported the findings and concluded that blinding had been effective.

In studies that are not blinded, the investigators must discuss the methods used to prevent bias. All relevant data must be presented if available, and the matter must be addressed in the discussion section (Table 1). The allocation sequence can be concealed even if the study is not blinded. That is, the person actually assigning the patient to a particular treatment does not know the order in which patients are to receive treatment, so bias from differential assignment (e.g., assigning sicker patients to the new treatment because the new treatment is believed to be better) need not occur even in a study that is not blinded. An effective method of concealing allocation is to require the person who actually assigns the patient to treatment groups to contact a research coordinator to obtain the assignment. That way, the person assigning the treatment does not have access to the allocation sequence.

Adverse effects from placebo administration. Closely related to blinding are adverse effects from placebo administration when placebos need to match certain characteristics (e.g., taste) of the test drug. In the study comparing the zinc lozenges with placebo, the placebo lozenges needed to be very similar to the zinc lozenges to maintain blinding.[17] The study authors described the placebo in the methods section and addressed the issue of adverse effects from the placebo (which would make the zinc lozenges appear effective) in the discussion section.

Adherence to the protocol. Adherence to the study protocol, the final internal-validity factor described in Table 1, can have a major impact on the interpretation of the findings. Consider the extreme, hypothetical case in which a significant difference is found to favor the treatment but the subjects in the treatment group do not take any of the medication. The observed effect could not be caused by the treatment if no one took it. Hence, adherence information is important to establishing the effects of treatment—and is considered an ethical imperative by some.[18]

Investigators need to be alert to all types of protocol violations. Both providers and patients may violate protocols. While the failure of patients to adhere to the protocol likely reduces the effect of the treatment, violations by providers and researchers may bias the study in either direction, depending on the particular violation.[10] For example, if data (e.g., serum glucose concentrations after an insulin dose) are collected at times different from those specified in the protocol, patients may display a different response than if the data were collected when they should have been. In that case, the effect of the treatment would appear more or less powerful than it really is.

Like study withdrawals and experimental mortality, nonadherence does not occur randomly. Factors that affect adherence, such as severity of illness, level of education, and socioeconomic status, may be independently related to treatment outcomes, so that responses in the adherent group are biased and not representative of the entire sample. Like other factors that may affect internal validity, adherence to the protocol and the standards of adherence used in the study are described in the methods section. When describing the findings relevant to adherence in the results section, the results are presented by treatment group so that any between-group differences in adherence are readily apparent. The results should include adherence problems associated with protocol violations by providers or researchers. If differences were identified, then the implications for interpreting the study findings should be discussed.

Scientific misconduct. Yet another problem related to establishing the effects of treatment is scientific misconduct. Fabrication of data and manipulation of data (such as discarding data that do not support the hypothesis) result, of course, in a study that has no internal validity. The findings cannot be replicated by other investigators, nor can there be generalization.

Methodological problems and statistical significance

Although problems with internal validity are typically associated with studies reporting statistically significant differences, methodological problems may introduce extraneous variance into the study that obscures the real differences and produces findings that are not significant.[19] (In the real world of research, controlling extraneous variance so that a real difference can be identified is the larger problem.) Below, we compare the implications of methodological problems both for results that are significant and for those that are not.

Experimental mortality. Experimental mortality may favor either

the treatment or the control group. If patients who are likely to improve anyway predominate in the treatment group through differential withdrawals and only the data from these patients are included in the analysis, then that group will appear to have better outcomes. If a similar scenario occurs for the control group, then the difference between the two groups may not appear to be significant. In addition, excessive withdrawals may reduce the sample size so that the power is no longer sufficient to detect a significant difference. In studies with high dropout rates in both treatment and control groups, both problems are likely operating, and the findings cannot be interpreted with any degree of confidence.

Randomization process. Significance may be affected by the randomization process if randomization results in unequal groups. Small groups (i.e., those with 100 or fewer subjects[3]) are especially vulnerable to unequal randomization effects. If the inequality favors the control group, then the difference between the groups may not be significant. In addition, bias may be introduced into the randomization procedure in certain circumstances. Bias may be a particular problem if the person interacting with the patient also makes the assignment and is not blinded to the allocation sequence. If the experimental treatment is seen as highly desirable or beneficial, the assignment may be biased so that the sickest patients are assigned to the treatment group. In that case, the control group may appear to have a better outcome.

Blinding. Lack of blinding can reduce the apparent effect of a treatment and result in statistically nonsignificant results. If subjects know that they are not receiving the treatment under study, they may make every effort to achieve the outcome anyway. Another potential problem arises when the control group is contaminated (i.e., receives at least some of the treatment). Studies comparing service options or general prevention programs may be particularly vulnerable to this problem because they cannot be blinded. For example, studies involving a reduction in smoking or a change in dietary habits often do not find differences between groups because the control group has adopted many of the behavioral changes that constituted the treatment.[20] In contrast, significance might be spuriously increased if persons collecting outcomes data are aware of treatment assignment, since they may rate the outcome for the treatment group more favorably.

Adherence to the protocol. Poor adherence to treatment protocols by participants can reduce the treatment impact and lead to differences that are not significant. Inadequate compliance reduces the power of a study so that larger samples sizes are required to identify significant differences. In some cases, the sample size may need to be increased by 50% to counteract a 20% reduction in drug adherence.[10] Treatment effects may also appear nonsignificant when subjects who are not likely to benefit from the therapy are included in the study. This again reduces the power of the trial so that a larger sample is required.

Nonadherence may also have implications for the applicability of the treatment: If subjects cannot adhere to the treatment regimen, then its usefulness is reduced. In a study of dietary fiber supplements for preventing colorectal adenomas, the authors discussed the possibility that subjects were unwilling to comply with the high-fiber regimen; the regimen may not have been a useful intervention.[15]

Establishing generalizability

When investigators think of generalizability, they typically think of extrapolating the results to other patient populations, depending on whether patients were selected for the study by means of random sampling techniques. Study results based on random samples are considered generalizable, while study results based on other methods of identifying patients are not. However, clinical studies rarely use random sampling techniques, because the identity of every eligible patient in the targeted population must be known at the beginning of the study for a random sample to be taken from it. Since clinicians cannot identify patients who will have a myocardial infarction, attempt suicide, or experience other clinical events that determine eligibility before the trial begins, random sampling of a population cannot be used. Also, random sampling does not guarantee generalizability. If the targeted population is a small subpopulation within a larger population, the results may not be generalizable to the larger population because it may not be adequately represented in the random sample. Other information is needed to establish generalizability.

Information for determining external validity is provided in the methods and results sections of a research report. In the methods section, inclusion and exclusion criteria help identify the population to which the results might apply. Additional information on generalizability is found in the data on demographic characteristics, diseases, and other characteristics of the study participants. By examining the characteristics of the study participants, readers can estimate if they are likely to obtain similar outcomes in their own patient population. For example, the results of a study that evaluates the efficacy of a specific treatment in elderly Caucasian men with coronary heart disease cannot be extrapolated to Hispanic women.

The report may include a statement describing the authors' assessment of the population to which the results can be generalized. For example, the authors of one study wrote, "The study population . . . was representative of patients 75 years of age or

younger who were not receiving long-term aspirin treatment and who had not recently undergone angioplasty or bypass surgery."[21] Alternatively, the population to which the results can (or cannot) be generalized may be described in the discussion of study limitations. For example, the investigators may state that the study was conducted in a primarily Hispanic population at a single practice site in the Southwest and that generalizability to other populations is unknown.

Steps in establishing internal and external validity

The three-step process shown in Table 2 can be used to assess the validity of a study's findings and determine if they are relevant to readers' practices. The first step in establishing validity is to assess the statistical conclusion. Only if the conclusion is valid is internal validity assessed; similarly, external validity is assessed only if internal validity is established. This is the decision process recommended by Campbell and Stanley[1] and Cook and Campbell.[2] If there is no significant difference among groups or the reader concludes the difference is not valid, there is no treatment effect and no cause-and-effect relationship to assess. One may want to examine threats to internal validity to determine if they may have introduced extraneous variance, but then the purpose of the assessment is no longer to determine if the findings are relevant to one's practice.

A similar logic exists with respect to external validity; if there is no internal validity, then there is no treatment effect to generalize. Hence, the question of generalizability becomes moot.

Discussion

The cognitive map presented offers a guide to addressing specific problems with the internal validity of experimental studies. This guide will help investigators structure the information required to establish a cause-and-effect relationship and will steer readers toward the same information as they assess validity. The clear delineation of specific threats to internal validity and of the relationships among sections differentiates the cognitive map from checklists and from more general structural approaches. Checklists are inventories of items that should be included in a research report.[12,22] Typically, they include many items addressing a broad range of issues, only some of which are specifically related to internal or external validity. In addition, the relationships among sections and the role of information are not readily apparent with checklists. The cognitive map described closely resembles the structure suggested in the Consolidated Standards of Reporting Trials (CONSORT) statement, which does focus on the key pieces of information needed to evaluate internal and external validity.[23,24] However, again, the cognitive map highlights the relationships and roles of information in the report, not just the content of each section.

Because checklists address a broad range of issues involved in reporting research, the cognitive map described should be seen as a supplement to checklists, not a replacement. Also, the map should be differentiated

Table 2.
Steps for Assessing Validity of an Experimental Study[a]

Step	Assessment Process	Decision
1. Validity of statistical conclusion	Assess statistical significance (i.e., *p* value is ≤0.05 and statistical results are valid).	Difference is real and is not likely due to chance variation; proceed to next step. OR Difference is likely due to chance variation; stop here.[b]
2. Internal validity	Assess internal validity on basis of research design and operational procedures.	Difference is most likely due to the treatment; proceed to next step. OR Difference is probably due to the effects of confounding factors or bias; stop here.
3. External validity	Examine inclusion and exclusion criteria and characteristics of study participants.	Study participants are similar to patients the report reader sees; the treatment should be useful. OR Study participants are very different from patients the report reader sees; the treatment may or may not be useful.

[a]Use these steps when determining if research findings are applicable to a particular practice situation.
[b]Internal validity may be assessed if the purpose is to determine if threats to internal validity may be producing extraneous variance that has obscured the treatment effect.

from checklists and scales used to assess the overall quality of a study. The purpose of checklists is to assess research that has been reported and not necessarily to assist investigators in structuring a report.[25]

The cognitive map is limited in that some knowledge of research design is required to adapt it to specific research situations. Also, while some aspects of the map, such as threats related to withdrawals, protocol adherence, and placebo use, can be adapted to other research designs, other designs have additional problems that must be considered.[7] Knowledge of statistical techniques, such as multivariate logistic regression, may be necessary to adequately address some questions about internal validity.

The cognitive map should improve pharmacists' ability to effectively communicate their research findings. Pharmacists who have conducted high-quality research can more accurately represent study quality in their report. Improvement of study reporting is a need that has been recognized in both pharmacy and medicine,[21,26] and structured approaches to writing are believed to help authors attend to essential details.[27] Indeed, an evaluation of the impact of the CONSORT statement found that journal articles were more likely to include checklist items after journals began using it.[28]

Conclusion

By using a cognitive map, investigators reporting an experimental study can systematically address internal and external validity so that the effects of the treatment are accurately portrayed and generalization of the findings is appropriate.

References

1. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin; 1963.
2. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Boston: Houghton Mifflin; 1979.
3. Elwood M. Critical appraisal of epidemiological studies and clinical trials. 2nd ed. Oxford, England: Oxford Univ. Press; 1998.
4. Schulz KF, Chalmers I, Hayes RJ et al. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA.* 1995; 273:408-12.
5. Moher D, Jadad AR, Nichol G et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials.* 1995; 16: 62-73.
6. Chalmers TC, Smith H Jr, Blackburn B et al. A method for assessing the quality of a randomized control trial. *Control Clin Trials.* 1981; 2:31-49.
7. Harrison DL, Draugalis JR. Critically evaluating research methods: an introduction. *Manag Care Med.* 1996; 3:23-7.
8. Rabow J, Charness MA, Kipperman JK et al. Learning through discussion. Thousand Oaks, CA: Sage; 1994.
9. Rosenwasser D, Stephen J. Writing analytically. Fort Worth, TX: Harcourt College; 1997.
10. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. 3rd ed. St. Louis: Mosby; 1996:204-22.
11. Everitt BS, Pickles A. Statistical aspects of the design and analysis of clinical trials. London: Imperial College Press; 1999.
12. Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. *JAMA.* 1994; 272: 1926-31.
13. Schulz KF, Chalmers I, Grimes DA et al. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA.* 1994; 272:125-8.
14. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet.* 1990; 335:149-53.
15. Alberts DS, Martinez ME, Roe DJ et al. Lack of effect of a high-fiber cereal supplement on the recurrence of colorectal adenomas. *N Engl J Med.* 2000; 342:1156-62.
16. Devereaux PJ, Manns BJ, Ghali WA et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. *JAMA.* 2001; 285: 2000-3.
17. Mossad SB, Macknin ML, Medendorp SV et al. Zinc gluconate lozenges for treating the common cold. *Ann Intern Med.* 1996; 125: 81-8.
18. Efron B. Foreword in special issue on analyzing non-compliance in clinical trials. *Stat Med.* 1998; 17:249-50.
19. Polk RE, Hepler CD. Controversies in antimicrobial therapy: critical analysis of clinical trials. *Am J Hosp Pharm.* 1986; 43:630-40.
20. Beresford SA, Curry SJ, Kristal AR et al. A dietary intervention in primary care practice: the eating patterns study. *Am J Public Health.* 1997; 87:610-6.
21. Theroux P, Ouimet H, McCans J et al. Aspirin, heparin, or both to treat acute unstable angina. *N Engl J Med.* 1988; 319:1105-11.
22. Asilomar Working Group. Checklist of information for inclusion in reports of clinical trials. *Ann Intern Med.* 1996; 124:741-3.
23. Begg C, Cho M, Eastwood S et al. Improving the quality of reporting of randomized controlled trials. *JAMA.* 1996; 276:637-9.
24. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *JAMA.* 2001; 285:1987-91.
25. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. *Int J Technol Assess Health Care.* 1996; 12:195-208.
26. Ferrill MJ, Norton LL, Blalock SJ. Determining the statistical knowledge of pharmacy practitioners: a survey and review of the literature. *Am J Pharm Educ.* 1999; 63:371-6.
27. Hartley J. From structured abstracts to structured articles: a modest proposal. *J Tech Writ Commun.* 1999; 29:255-70.
28. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials. *JAMA.* 2001; 285: 1992-5.