Internal and External Validity in Two Studies That Compared Treatment Methods

Jeffrey A. Gliner

Key Word: research design

Research comparing the effectiveness of two treatments offers both strengths and weaknesses for occupational therapy. Although it is worthwhile to determine which of two treatments works best for a particular problem, methodological problems may arise that preclude a valid conclusion. To draw valid conclusions from research, criteria for internal validity and external validity must be satisfied.

The two preceding articles in this issue are examples of studies that used between-groups experimental methodology to compare the effectiveness of two different treatments. This paper evaluates the above-mentioned studies on the basis of principles of internal and external validity. One of these studies (Jongbloed, Stacey, & Brighton, 1989) was truly experimental, whereas the other study (Groves & Rider, 1989) was quasi-experimental. Results from both studies were similar because, in each study, both of the treatment groups improved, but there were no significant differences between treatments. Absence of a true control group in both studies presented limitations on the conclusion that both treatments worked equally well.
ternal validity, statistical conclusion validity, and construct validity of causes or effects—which, when violated, could lead to invalid conclusions.

**Internal validity** usually refers to the “tightness” or control of the design; it allows the conclusion that changes in the dependent (measured) variable were in fact due to manipulation of the independent variable (usually considered the variable of interest). Campbell and Stanley (1966) considered internal validity to be basic to the interpretation of any experiment.

**External validity** refers to generalizability of the results of the study to other populations, treatments, or settings. Unfortunately, it is often the case, especially in applied fields, that as the design becomes more controlled, generalizability is sacrificed, or as generalizability is increased, good design is sacrificed. However, as Campbell and Stanley (1966, p. 5) pointed out, “While internal validity is the *sine qua non*, and while the question of *external validity*, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal.”

**Statistical conclusion validity and construct validity of causes or effects** are subsets of internal and external validity. Statistical conclusion validity refers to the proper selection of statistical procedures, whereas construct validity of causes or effects refers to the generalizability of the independent and dependent variables toward theoretical constructs. This article will consider only internal and external validity.

For a more detailed account of statistical conclusion validity for pretest-posttest designs, see Huck and McLean (1975) or Reichardt (1979).

Unfortunately, the clinician often does not have the opportunity to design studies that will be strong on both internal and external validity. However, internal and external validity are not all-or-none criteria, and studies may be carried out that have fairly good validity. The point is that the researcher must be aware that a study with problems in internal or external validity has limitations, and the study conclusions must be stated within these limitations.

The above criteria for a good study design can be used to consider the two previous papers and how they satisfied the conditions of internal and external validity.

**Internal Validity**

To conclude that the manipulation of the independent variable caused changes in the dependent variable, certain threats must be ruled out. These threats, or confounding factors, occur when the researcher fails to control some extraneous variable (Cozby, 1985); in such a situation, the researcher cannot be sure that the result was due only to changes in the independent variable. There are two general characteristics of sound experimental research (Cozby): manipulation of the independent variable and control of extraneous variables. Control of extraneous variables can be obtained by (a) randomly assigning subjects to groups and (b) having a control group or groups that have all of the same features as the experimental group except for the independent variable. Although there are numerous threats to internal validity (Campbell & Stanley, 1966), only those that appeared to create problems in interpretation for either of the preceding studies were discussed.

**Manipulation of the Independent Variable**

Both studies satisfied the first criterion of a true experiment, in that an independent variable was manipulated (i.e., that subjects could be assigned to either condition of the independent variable). This differs from a study with a nonmanipulated independent variable or attribute variable (Kerlinger, 1973), such as the site of a lesion in a cerebrovascular accident, where subjects cannot be assigned to conditions. The independent variable in the Jongbloed et al. study was treatment, of which there were two levels, functional treatment and sensorimotor integrative treatment. The independent variable in the Groves and Rider study was postoperative treatment, of which there were also two levels, progressive resistive exercise and no structural exercise.

**Control of Extraneous Variables**

Random assignment of subjects to groups. Jongbloed et al. satisfied this characteristic of a research study in that they were able to randomly assign subjects to groups. Random assignment removes the internal validity threat of selection (Cook & Campbell, 1979). This strengthened the study and distinguished it from many other treatment comparison studies in the rehabilitation medicine literature. (Ottenbacher, 1982, found that of the eight studies with similar types of methodology reviewed, only half randomly assigned subjects to groups.) Groves and Rider on the other hand, could not randomly assign subjects to groups. As mentioned by the authors “... it would be unethical for [the clinic representatives] to implement a program that was not used commonly in their clinic” (p. 401). This weakened the design considerably by introducing *selection differences*, described by Cozby (1985) as occurring “when subjects who form the two groups in the experiment are chosen from existing natural groups” (p. 58). The design for the Groves and Rider study would be appropriately called a non-equivalent control group design with a pretest and a posttest (Cook & Campbell, 1979; Kenny, 1975). It
should be noted that the term *equivalence* in this case does not refer strictly to the equivalence of the dependent measures recorded during the pretest (which, unfortunately, the authors failed to provide). Even if the dependent measures, such as grip strength, were not different as determined statistically during the prephase of the experiment, other variables could have influenced the outcome of the study due to self-selection of the subjects in a particular clinic.

**Proper control group.** A sound research study includes a proper control group. In the strictest sense, the control group has all of the characteristics of the experimental group except the independent variable. However, the term *control group* can lend itself to different interpretations, all of which depend on the question that is asked. The purpose of both studies was to determine which of two different treatment conditions worked best for a particular disability condition. The authors of both studies did not appear to be specifically interested, at least at the time of their writing, in why either treatment worked, and therefore the two treatments were largely different. Furthermore, the ethics of having a particular nontreatment control group (Ottenbacher, 1986) has been addressed and remains a problem for clinical disciplines.

The problem of not having a control group that received no treatment was illustrated in the Jongbloed et al. study. Two treatments were compared over either two or three time periods (admission and 8 weeks or admission, 4 weeks, and 8 weeks). At the end of 8 weeks, all but three of the outcome measures were found to have changed significantly, as determined by the appropriate statistical test. However, for the most part, there were no statistically significant differences between treatment groups. Jongbloed et al., after pointing out the limitations of their study, concluded, "occupational therapists can consider using either approach in planning treatment for CVA patients" (p. 395). This conclusion implied that both treatments worked equally well. However, from the design of the study, this conclusion was not justified. The authors even pointed out that the role of spontaneous recovery could not be assessed. The problem was the internal validity threat of maturation; subjects in both of the treatment groups could have recovered spontaneously (over time), and this was not the variable of interest.

Furthermore, a potential confounding variable was that both groups received physical therapy. Since there were no differences between groups on the outcome measures, any increases could have been attributed to physical therapy. This problem concerns the internal validity threat of history, because another variable or event that was not of primary interest to the study would have interacted with the independent variable. Other studies in the rehabilitation medicine literature also have not used a nontreatment control group. However, as Basmajian et al. (1987) pointed out, "Our resolve to compare two active therapies and not to use 'do nothing' controls was based on evidence supporting EMGBF [electromyogram biofeedback] as an effective therapeutic adjunct for the hemiplegic upper limb." Jongbloed et al. did not provide similar types of support in their article.

The study by Groves and Rider employed a control group that was closer to the "do nothing" group than was that of Jongbloed et al. However, it was unclear whether the warm soaks and performance of daily tasks were also encouraged in the experimental group. The conclusion reached by Groves and Rider, which questioned "the efficacy of providing costly time-intensive postoperative treatment if routine activity and warm soaks are equally effective" (p. 401), may have been misleading. Just as in the Jongbloed et al. study, no group was included that did absolutely nothing, so that one might question the healing properties of warm soaks and routine activities. Also, some statistical problems raised the question of whether either group improved. The authors did not explain whether they used a gain-score analysis or a posttest-only analysis for the statistical treatment of their data (Kenny, 1975; Reichardt, 1979). With either analysis, given no statistically significant differences between groups, it would have been impossible to determine whether either group improved.

**Testing**
One other threat to internal validity not mentioned above is testing (Cook & Campbell, 1979). The threat of testing occurs when the same test is given more than once (i.e., pretest and posttest) and, therefore, the respondent's familiarity with the test could bias the results. This threat is often difficult to assess, but I think that the type of measures employed in both studies were not apt to be influenced by pretesting.

Another threat of testing is bias by the investigator judging performance. Jongbloed et al. used a double-blind technique to solve this problem, neither the evaluator nor the subjects knew which treatment group they were involved with. Groves and Rider pointed out in their study that the recorders were aware of the treatment condition; thus, bias could not be ruled out.

**External Validity**
External validity refers to generalizability of results to populations, settings, and times (Cook & Campbell, 1979). Problems in external validity are primarily problems in selection. Specifically, a clinician would...
question whether subjects, treatments, and measures were selected in such a way as to be generalizable to a larger group. External validity is important because the purpose of research is usually to extend findings beyond the specific sample of the experiment. However, as stated earlier, the question of external validity is one of induction and is never completely answerable (Campbell & Stanley, 1966). Bracht and Glass (1968) divided external validity into population validity (generalizable to populations) and ecological validity, which asks, "Under what conditions, i.e., settings, treatments, experimenters, dependent variables, etc., can the same results be expected?" (p. 438). From the above statement, numerous threats to external validity could result. This paper addresses three major threats to external validity that were important in clinical studies: sample, treatment, and measures representations.

**Representation of the Sample**

The threat to external validity is selection of the sample. Was the sample randomly selected from a larger population of persons with similar disabilities? Both studies appeared to have used convenience samples, because inclusion in the study necessitated that subjects be patients in a particular hospital or clinic. An important question to ask here is, "Were the cerebrovascular accident patients or carpal tunnel syndrome patients representative of their respective populations?" Under the strictest of criteria, the answer would be no.

However, external validity is not an all-or-nothing phenomenon, and common sense may also play a role. For example, the question could be asked, "Is there any reason that this particular group of patients should be different or less representative of the general population of this disability?" When considering the purpose of the study, one could ask, "Is there any reason to suspect that this group should respond differently to the treatment than would any other sample from this population?" Furthermore, the Jongbloed et al. study included 90 patients. The Groves and Rider study included a smaller number of patients but could still be considered representative of the population of carpal tunnel syndrome patients.

Another feature of patient representation is heterogeneity of the sample. According to their results section, Jongbloed et al. appeared to use a relatively homogenous sample of cerebrovascular accident patients. Groves and Rider demonstrated a relatively heterogenous age range, with mostly female subjects, but other aspects of the sample were not given.

**Treatment Representation**

A second criterion for external validity is the representativeness of the treatment regimen. Would a similar study that used the same therapy (e.g., sensory integration) use a similar protocol? Jongbloed et al. provided previous documentation for both types of treatment used in their study, although it was unclear whether other studies had used the same treatment regimen. Groves and Rider, however, pointed out that there were no published articles evaluating treatment of carpal tunnel syndrome at the 3-week postoperative stage. Both types of treatments were documented only through personal communications. Strengths for both studies were that they were using regular therapeutic techniques applied at their respective clinic or hospital with experienced personnel.

Length of treatment could be considered another characteristic of treatment representation. Groves and Rider suggested that a longer treatment period, as demonstrated by Trombly (1983), might have yielded greater change. Jongbloed et al. used an 8-week treatment period, but it was not clear whether this was a representative length for both treatment approaches.

**Representation of Measures**

Both studies used measures that attempted to assess success of the treatment. This criterion for external validity could be interpreted in two ways. First, the measures should be representative of similar types of studies attempting to assess similar therapeutic interventions. Second, and perhaps more important, the measures of any clinical research study should be clinically significant (Garfield, 1978; Lick, 1973). Jongbloed et al. combined measures that were representative of sensorimotor integration and functional therapy. Further, the study attempted to choose measures that would reflect functional gains, such as the Barthel Index and meal preparation. Groves and Rider used measures of wrist flexion, wrist extension, and grip strength. All of these measures appeared to have been used in both clinics and had been mentioned as useful goals by an occupational therapist in the literature review. The authors did not say how large the changes in these measures needed to be to represent functional gains.

**Conclusion**

Internal and external validity were assessed on two studies that used a research methodology characterized by large-N group comparisons. This type of design has advantages over single-subject designs, considering the general question of assessment of differences between two treatment methods. Although the two studies examined different methods used to treat different handicapping conditions, the general design for both studies was considered to be a pretest-posttest control group design. However, Jongbloed et al. randomly assigned subjects to groups (a true experi-
ment), whereas Groves and Rider used already selected groups (a quasi-experiment). Kenny (1975, p. 360) summed up the situation as follows: "The difference between the true experiment and the quasi-experiment is of the magnitude of the difference between sight and blindness. We must often grope in the darkness with quasi-experimental designs, but this blindness both forces us to compensate for biases and helps us develop a newfound sensitivity to the structure of data. Finally, it makes us appreciate the clarity of true experimental inference."

References


