Record: 1

Title: Threats to the Validity of Research.

Author(s): Parker, Randall M.

Source: Rehabilitation Counseling Bulletin, v36 n3 p130-38 Mar 1993.

Availability: Y

Peer Reviewed: 0034-3552

ISSN: Research Design, Research Problems, Validity

Descriptors: Notes that, when planning research, obvious strategy is to assess likely

Abstract: flaws and develop approaches to overcome them. Introduces set of concepts that researchers can use to evaluate quality of their designs. Presents 32 threats to validity organized according to 4 types of validity (internal, external, statistical conclusion, and construct validity) described by Cook, Campbell, and Peracchio (1990). (NB)

Number of References: 0

Language: English

Publication Type: Journal Articles; Opinion Papers

Journal Code: CIJJUL1993

Entry Date: 1993

Accession Number: EJ458938

Database: ERIC

THREATS TO THE VALIDITY OF RESEARCH

About 2 1/2 years ago I was asked by J. Lee Wiederholt, the editor of the Journal of Learning Disabilities, to review three studies and comment on their relative methodological merits and flaws (Parker, 1990). The studies all attempted to assess the effects of individually prescribed, tinted lenses (called Irlen lenses after the founder, Helen Irlen) on the reading performance of children with dyslexia and other reading difficulties. The approach I took was to review the articles using the concepts provided by Cook and Campbell (1979, 1983). These concepts, which are referred to as threats to the validity of research, may be viewed as tools with which to ferret out weaknesses in research designs. With permission of the JLD editor, I present here an updated description of these threats, based largely on a more recent publication by Cook, Campbell, and Peracchio (1990).

It is a widely accepted truism that all published research is flawed to some extent. Because the research enterprise is fraught with many pitfalls, researchers must become well versed in recognizing and, when possible, avoiding design shortcomings. An obvious strategy when planning research is to assess the likely flaws and develop approaches to overcome them. To accomplish this assessment, researchers require a set of concepts to evaluate the quality of their designs. Such a set of concepts has been detailed extensively in the literature (Campbell & Stanley, 1966; Cook & Campbell, 1979, 1983; Cook, Campbell, & Peracchio, 1990). Cook, Campbell, and Peracchio (1990) presented 32 threats organized according to four types of validity, namely internal, external, statistical conclusion, and construct validity. The purpose of this editorial is to describe the four types of research validity and the 32 threats to research validity.

INTERNAL VALIDITY

Internal validity, the most important type of research validity, refers to the extent to which extraneous variables (that is, sources of error variance) are controlled. Failure to control extraneous variables prevents the researcher from concluding that observed outcomes are due to the independent variable(s). Experimental control may be attained by using several techniques, including (a) assigning participants randomly to treatment and control groups, (b) holding extraneous variables constant or restricting their range, (c) including extraneous variables in the design to isolate their effects, (d) employing methods of statistical control (e.g., analysis of covariance), and (e) matching participants in the treatment and control groups on contaminating, extraneous variables (Bolton & Parker, 1992; Christensen, 1980). Failure to use adequate methods to control error variance results in nine possible threats to the internal validity of the research.

Threats to Internal Validity

Threats to internal validity include the following:

1. History. History refers to an extraneous event that correlates with the dependent variable and occurs during the study. It may spuriously mask or enhance the outcome variable. History would threaten the validity of a study on the relationship of stressful instructions to test anxiety if several experimental group participants received IRS audit notices just before the posttest was administered. Their heightened stress and anxiety might spuriously raise the test anxiety scores of the experimental group, leading to a questionable conclusion that the instructions alone raised test anxiety levels.
2. Maturation. This threat refers to uncontrolled, naturally occurring, developmental changes in research participants that affect their performance on the outcome variable. For instance, during a 2-year study of the effects of a new instructional method on the reading skills of elementary students, the treatment group (receiving the new method) outperformed the control group (receiving the traditional method). Upon further study, however, it was discovered that the treatment group had a higher proportion of female students than did the control group. Furthermore, the girls' reading achievement grew at a faster rate than the boys'. Finally, it was found that the performance differential between the experimental and control groups could be explained by maturational differences in reading skills between boys and girls.
3. Testing. Pretesting may sensitize participants in ways that affect posttest scores. For example, students may remember their pretest responses and answer more items correctly at posttest. On occasion the opposite result occurs. Participants may react negatively to taking the same test twice, and on posttest answer items haphazardly and obtain a spuriously low score.
4. Instrumentation. This refers to deterioration or changes in the accuracy of instruments, devices, or observers used to measure the dependent (outcome) variable. Examples of instrumentation include observers forgetting their training and beginning to record observations inaccurately, and surveys mailed to one geographical area becoming wrinkled and rain soaked in transit, affecting both legibility of the surveys and accuracy of the responses.
5. Statistical regression. Grouping participants on the basis of extreme scores may result in inaccurate categorizations, because extreme scorers tend to regress toward the group mean on repeated testing. For example, high and low reading groups are formed by dividing students at the mean on a reading test. To the extent that the reading test is unreliable, individuals who scored the highest will tend to score lower, and those who scored lowest will tend to score higher at posttest. Furthermore, some students scoring slightly above the mean at pretest will score below the mean at posttest, and the reverse is true for those scoring slightly below the mean. The migration of participants across the high-low group boundary due solely to test unreliability results in misclassification and inflated error variance.

6. Selection. This threat occurs when participants volunteer for a treatment or are assigned to treatment and control groups based on their preferences. This assignment results in the groups' being different on many variables. Consider the fictitious example of a study conducted to identify client characteristics associated with positive outcomes in individual versus group counseling. If participants were allowed to choose the counseling approach, volunteers who self-selected into individual counseling may differ from those selecting group counseling in ways that affect the outcome measure.
7. Mortality. This threat refers to the loss of participants and their data during the course of a study due to illness, forgetfulness, death, or other causes. For example, when the treatment itself causes participants to drop out of the study, the treatment groups' posttest mean would be contaminated because only the survivors' scores would be available.
8. Interactions with selection. Many of the foregoing threats to internal validity may interact with selection to produce effects that may- be erroneously attributed to the treatment. People who volunteer for a treatment might be different from a nonvolunteer control group on variables that are related to the outcome variable. Interaction of selection and maturation, for example, might lead to the treatment and control groups' being composed of individuals who are maturing at different rates; this difference might contaminate measures of the outcome variable.
9. Ambiguity about the direction of causal influence. This threat to internal validity is common in bivariate correlational studies when it is not clear whether A causes B or B causes A. Consider a situation in which supervision and productivity are negatively related. Low supervision is related to high productivity and vice versa. Does low supervision cause high productivity, or does high productivity cause low supervision?

EXTERNAL VALIDITY

External validity refers to the degree to which research findings can be generalized across persons, times, and settings. Generalizing across persons requires research samples to be representative of the population of interest. Generalizing across times and settings typically necessitates systematically administering the experimental procedures at different times and in different settings. Failing to obtain a sample that is representative of a population, of a range of settings, and of a range of times results in the inability to generalize the findings of the study beyond the persons, setting, and time employed in the research. Such failures pose serious threats to external validity.

Threats to External Validity

Specific threats to external validity include the following:

- 10. Interaction of treatments with treatments. When multiple treatments are administered to the same participants, the effects may be cumulative. Thus, the results of research employing multiple treatments may not be generalizable to situations in which the treatments are given singly. This threat is particularly applicable to time-series designs involving two or more treatments.
- 11. Interaction of testing with treatment. The pretest may increase or decrease the respondents' responsiveness or sensitivity to the treatment. As a consequence, the results are not generalizable to the nonpretested population from which the treatment group was selected.
- 12. Interaction of selection with treatment. This threat occurs when research participants are volunteers, that is, individuals who are prone to seek out research participation. Such persons may have traits that tend to enhance or diminish the effects of the treatment. Thus, the results are not generalizable to a population that includes nonvolunteers.

- 13. Interaction of setting with treatment. Treatments demonstrated in one environment, for example, the laboratory, may not work in other settings, for example, the. classroom. Therefore, this threat to external validity refers to whether effects demonstrated in one setting are generalizable to other settings.
- 14. Interaction of history with treatment. The effects observed in a study may be due to special circumstances, for instance, teachers showing a substantial lowering of anxiety after a stress reduction program. During the study, however, a tornado warning was sounded and shortly thereafter canceled. Is the observed effect generalizable to other, more normal circumstances?

STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity is concerned with the appropriate use of statistics to arrive at accurate decisions about accepting or rejecting hypotheses. Before proceeding, a few terms must be defined. Let us consider a typical study in which treatment and control groups' performances are being compared on an outcome variable. The null hypothesis states there is no difference between group means on the outcome variable. In this situation the Type I error, which is called alpha (a), is the probability of falsely rejecting the null hypothesis, that is, finding a statistically significant difference when the means come from the same population. On the other hand, the Type II error, beta (Beta), is the probability of failing to reject a false null hypothesis (i.e., finding no statistically significant difference when the means come from different populations).

Threats to Statistical Conclusion Validity

Threats to statistical conclusion validity include the following:

- 15. Low statistical power. Statistical power is equal to the quantity 1.0 - Beta, or the probability of rejecting a false null hypothesis (that is, finding a statistically significant difference when the sample means, in fact, come from different populations). Power is a function of alpha, sample size (N), and effect size (ES). Effect size (ES) refers to the amount of common variance between the independent variable (IV) and the dependent variable (DV), or the degree to which changes in the IV result in changes in the DV. Increasing the alpha level, sample size, and effect size, singly or in combination, increases statistical power. According to Cohen (1988), power ideally should be about .80. That is, researchers should design their studies so that they have 8 in 10 chances of obtaining a statistically significant result in the sample data when one actually exists in the population data. Surveys of research, however, indicate that most studies have much lower power (Cohen, 1962; Lipsey, 1990). Through conducting power analyses while designing a study and systematically varying a, N, and ES, the researcher can approximate the desired level of power in the resultant research (see Szymanski & Parker, 1992).
- 16. Fishing and the error rate problem. Running many statistical tests on one data set is called fishing. Fishing produces error rates that are higher than the preset alpha. Running one statistical test on a data set results in a Type I error rate equal to the preset alpha (usually .05). Running two or more statistical tests, however may inflate alpha above the predetermined rate. (Exceptions include follow-up tests, for example, Scheffe's, Tukey's, and similar multiple comparison tests, as part of an ANOVA or ANCOVA design.) Various corrections for alpha inflation are possible including Scheffe's, Tukey's, and similar multiple comparison tests as part of an ANOVA or ANCOVA design. A more general procedure is called the Bonferroni correction, which is accomplished by setting alpha equal to the desired alpha (e.g., .05) divided by the number of tests run for all hypotheses or, alternatively, the number of tests for a logically related subset of hypotheses (see Parker & Szymanski, 1992).
- 17. Low reliability of measures. Measures with low reliability increase error variance and reduce the power of statistical tests. For instance, decreased reliability often occurs when simple gain scores (posttest minus

pretest scores) are used as a measure of change on the dependent variable. Change scores are notoriously unreliable, because by subtracting the pretest from the posttest, one is left with less true variance and relatively more error variance. Residual gain scores (posttest scores with pretest covariance partialled out) frequently are preferable to simple gain scores. In general, the reliability of measures may be enhanced by adding more items with high internal consistency, decreasing the intervals between pretests and posttests, and using corrections for unreliability (e.g., correction for attenuation) with great caution.

- 18. Low reliability of treatment implementation. This threat is due to the lack of standardization of procedures used to administer the treatment. Using different individuals and different occasions to implement the treatment increases this threat, as does the failure to adequately train individuals administering the treatment.
- 19. Random irrelevancies in the experimental setting. Scores on the outcome variable may be affected by aspects of the experimental setting other than the treatment. This threat may be reduced by choosing settings free of distractions (e.g., irregularly occurring noise) or by focusing participant attention on the treatment and lowering the salience of extraneous environmental factors.
- 20. Random heterogeneity of respondents. This threat occurs when respondents are heterogeneous on variables related to the outcome variable. When respondent variables correlate with the outcome variable, error variance is inflated. This problem may be controlled by selecting participants who are homogeneous on all variables related to the dependent variable, excluding the independent variable. Alternatively, participants may be "blocked" on such variables, which would then be included in the statistical design.

CONSTRUCT VALIDITY

The construct validity of a variable refers to whether the variable is adequately defined and accurately measured by the instruments, procedures, manipulations, and methods employed in the study. A valid construct must have a unique operational definition. When a construct is suggested as a cause or effect, it is valid only when other constructs cannot be construed as being the cause or the effect.

Threats to Construct Validity

Threats to construct validity include the following:

- 21. Inadequate preoperational explication of constructs. When constructs are poorly defined initially, the instruments, procedures, manipulations, and methods used in the study cannot be adequately specified and may bear little relationship to the constructs being studied. Consequently, the results of the study cannot be accurately attributed to the constructs of interest.
- 22. Mono-operation bias. Construct validity is limited when a construct is defined by only one measure or operation. Multiply operationalized variables tend to be more valid than single operations because single operations underrepresent constructs and contain irrelevancies. Alternative measures of a target allow one to triangulate on the construct.
- 23. Monomethod bias. This threat occurs when all the manipulations and measures use the same means of presenting the treatments or recording the results. For example, because many leadership studies have employed single paper-and-pencil measures, it has been suggested that leadership theories predict paper-and-pencil behavior better than actually practiced leadership behaviors (Campbell, Daft, & Hulin, 1982). Other measures, such as observations, should have been used in these studies.
- 24. Interaction of procedures and treatments. Participants who receive new information or have new experiences as part of the treatment may react differently to the treatment. For example, participants in a 2-

month study using tinted glasses to improve their reading performance may react to the fact that at the termination of the study they would be allowed to keep the glasses. If an improvement in reading is observed, is it due to the effect of the tinted lenses, motivation caused by giving the glasses as a gift, or a combination of both? The latter case is an example of an interaction between a procedure (giving a gift) and the treatment (effect of tinted lenses).

- 25. Diffusion or imitation of the treatment. If the treatment is widely disseminated or the treatment and control groups communicate about the treatment, the control groups may seek out and receive the treatment. In this case the groups may become indistinguishable on the dependent (outcome) variable.
- 26. Compensatory equalization of treatment. When the treatment is desirable, research administrators may be reluctant to tolerate this inequity. In an educational research project, for example, administrators gave needy control schools funds in amounts equivalent to the experimental schools in contradiction to directions. This action obviously introduced contamination into the study.
- 27. Compensatory rivalry. When participants' membership in experimental or control groups is made public, competitive motivations may result. Saretsky (1972) noted that the performance of students taught by control teachers in the office of Equal opportunity Performance Contracting Experiment was higher during the experiment than in previous years. Apparently, performance contracting was perceived by the teachers as a threat to their job security, causing them to redouble their teaching efforts. Saretsky called this the "John Henry effect," comparing the control group's extra efforts to those of John Henry, a folk song character who competed with a machine in laying rail for the railroad.
- 28. Resentful demoralization of respondents receiving less desirable treatments. Particularly when an experiment is obtrusive and the experimental group receives preferential or desirable treatment, the no-treatment control group may feel resentful and demoralized. This resentment may lead to posttest differences that are not due to the positive effects of obtaining the treatment on the experimental group, but to the negative effects on the control group of not receiving the treatment.
- 29. Hypothesis guessing within experimental conditions. Research participants receiving a treatment may attempt to guess the purpose of the study and alter their outcome behavior accordingly. The well-known Hawthorne effect is the result of hypothesis-guessing by experimental participants. Hypothesis guessing may be avoided by making the hypotheses difficult to guess, by reducing the reactivity and obtrusiveness of the study, and by purposefully giving different bogus hypotheses to different participants. The latter action, however, may lead to ethical violations.
- 30. Evaluation apprehension. Treatment group participants, because of anxiety about being evaluated by the researcher, may attempt to present themselves in a favorable light. If this favorable presentation affects the outcome variable, the experimental results will likely be contaminated.
- 31. Experimenter expectancies. Research suggests that experimenter expectancies may influence participants' behavior and the outcome data. This effect may be avoided by selecting objective individuals who do not know the purposes of the study to administer the treatments and record the data.
- 32. Confounding levels of constructs and constructs. This threat occurs when the independent variable (IV) has multiple levels. One may conclude, for example, that the overall effect of an IV with four levels (e.g., four levels of drug dosage) is zero when only IV1 (level one of the IV) and IV4 do not affect the dependent variable (DV). IV2 and IV3, however, might affect the DV but remain undetected because of an overall nonlinear relationship. This threat typically occurs with weak treatments. Conclusions, nonetheless, are drawn regarding the ineffectiveness of the treatment without noting the nonlinear relationship or weakness

of the manipulation. This threat can be controlled by manipulating many levels of the IV and measuring many levels of the DV.

CONCLUSION

I suggest researchers use their knowledge of the 32 threats in planning research. Although no study is flawless, they can use this information to improve their designs. In addition, researchers should describe all major validity threats in research manuscripts. Authors have a responsibility to acknowledge the weaknesses of their research and inform readers regarding the degree of care that must be taken in interpreting and generalizing the results. I hope that this editorial will assist authors to that end.

Editor's Note. This article is adapted from: "Power, control, and validity in research" by Randall M. Parker, 1990, Journal of Learning Disabilities, 23, 613-620. Copyright by PRO-ED, Inc. Reprinted with permission. The author thanks Edna Szymanski for her comments and suggestions on an earlier draft of this editorial.

REFERENCES

Bolton, B., & Parker, R. (1992). Research in rehabilitation counseling. In R. Parker & E. Szymanski (Eds.), Rehabilitation counseling: Basics and beyond (2nd ea., pp. 333-364). Austin, TX: PRO-ED.

Campbell, J., Daft, R., & Hulin, C. (1982). What to study: Generating and developing research questions. Beverly Hills, CA: Sage.

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

Christensen, L. (1980). Experimental methodology. Boston: Allyn & Bacon.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (3rd ed.). New York: Academic Press.

Cook, T., & Campbell, D. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Cook, T., & Campbell, D. (1983). The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 223-326). Chicago: Rand McNally.

Cook, T., Campbell, D., & Peracchio, L. (1990). Quasi experimentation. In M. Dunnette & L. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ea., Vol. 1, pp. 491-576). Palo Alto, CA: Consulting Psychologists Press.

Lipsey, M. (1990). Design sensitivity: Statistical power for experimental research. Newbury Park, CA: Sage.

Parker, R. (1990). Power, control, and validity in research. Journal of Learning Disabilities, 23, 613 620.

Parker, R., & Szymanski E. (1992). Fishing and error rate problem. Rehabilitation Counseling Bulletin, 36, 66-69.

Saretsky, G. (1972). The OEO P.C. experiment and the John Henry effect. Phi Delta Kappan, 53, 579-581.

Szymanski, E., & Parker, R. (1992). Statistical power in rehabilitation research. Rehabilitation Counseling Bulletin, 36, 2-5.

~~~~~~~~

Randall M. Parker Editor
~~~~~~~~