Epistemology and Experimentation: An Examination of Quality Factors in Research Design

Kenneth J. Ottenbacher

Key Words: data collection • research design

Forty-four research articles published in the occupational therapy literature were examined to determine the effect of design quality on study outcome. All 44 of the studies examined involved a parallel-groups comparison design. Twenty-two of the studies included random assignment of subjects to various groups, and the remaining 22 investigations used some nonrandom method to determine subject allocation. A standardized metric (i.e., effect size) was used to determine the effect of the independent variable in the 44 studies. The data analysis revealed that effect-size values were not significantly different between studies that involved random assignment and those not involving random assignment. The argument is made that such design characteristics as random assignment should be examined as moderator variables in any attempt to synthesize findings from multiple studies. Such an approach would treat the design used in a study as one of many possible variables that could influence the outcome. This approach would modify the a priori assumption that one research design is inherently superior to another regardless of the research question or context.

In discussing the generation of knowledge through traditional experimental and quasi-experimental approaches, Cook and Campbell (1979) observed that "the epistemology of causation, and of the scientific method more generally, is at present in a productive state of near chaos" (p. 10). The productive chaos referred to by Cook and Campbell is clearly evident in the current discussions of research approaches appearing in the occupational therapy literature (Christiansen, 1986; Gibson, 1984; Hasselkus, 1991; Mosey, 1989; Yerxa, 1987, 1988). Several of these articles have expressed dissatisfaction with conventional research approaches and argued that alternative methods are needed in applied fields such as occupational therapy. For example, Yerxa (1991) recently noted that "we as occupational therapists need to seek or invent new ways of knowing. These ways will need to be of a different order than those of the physical sciences (i.e., experiments that use statistical tools to determine significance are the methods of choice in the physical sciences)" (p. 200). Yerxa went on to state that "these new ways of knowing may include what are now called qualitative, naturalistic, or new paradigm research approaches" (p. 200).

The dissatisfaction with traditional experimental procedures in applied fields such as occupational therapy is both philosophical and practical. True experimental designs are often difficult or impossible to implement in applied environments. The quasi-experiments that are possible are frequently criticized as being scientifically uninterpretable and of poor research quality.

The Quality of Quantitative Research Approaches

Despite the philosophical and practical difficulties associated with traditional empirical approaches, the experimental model continues to be enthusiastically supported in many rehabilitation-related fields. In discussing the application of various empirical approaches in rehabilitation research, Reilly and Findley (1989) noted that "true random experiments are the best way to establish causation" (p. 197). The true experimental design referred to by Reilly and Findley and described by Campbell and Stanley (1963) in their classic monograph involves a comparison in which participants are randomly assigned to groups or conditions. This random assignment provides an equality between groups based on probability. That is, the probability is high that the two groups are equal on the outcome measure of interest. The subjects in one group are exposed to the independent, or treatment, variable, and the performance of both groups is measured on some operationally defined dependent (outcome) variable. Based on the obtained results, a null hypothesis, which generally states that there is no difference between the performance of the two groups, is evaluated. If the results are found to be statistically significant...
at a predetermined probability level, then the null hypothesis is rejected. The alternate hypothesis, that the two groups differ on some performance characteristics, is supported.

The above framework, schematically depicted in Figure 1, is often referred to as a randomized controlled trial in the biomedical literature or a true experimental design in the behavioral and social sciences. Properly implemented, the design provides convincing evidence regarding a cause–effect relationship between independent and dependent variables. The primary advantage of the true experimental design is that various threats to the internal validity of the study conclusions are reduced or eliminated (Campbell & Stanley, 1963; Cook & Campbell, 1979).

The internal validity of a study relates to the degree of confidence that the investigator (and reader) can have that any change in performance observed in the treatment group is a function of the independent variable and not of some extraneous or uncontrolled factor. For example, if no comparison or control group is included in a study, changes in performance of the subjects in the treatment group may have occurred naturally due to maturation, spontaneous recovery, or some factor other than the treatment. Campbell and Stanley (1963) stated that "internal validity is the basic minimum without which any experiment is uninterpretable" (p. 5).

The disadvantage of the true experimental configuration is that it is often difficult or impossible to implement in applied settings. The limitations of true experiments, or randomized controlled trials, in clinical environments were discussed by Kramer and Shapiro (1984), who noted that "despite the obvious advantages and impressive track record of RCT's [randomized controlled trials], clinical investigators have become increasingly aware of certain difficulties in their interpretation, feasibility, and ethics" (p. 2739). The difficulty of meeting the requirements of true experiments in clinical settings has forced investigators to use modified versions of classic experimental designs. Variations of the true experimental design are considered quasi-experiments. They do not provide the same degree of methodological control or inferential confidence as true experiments. The use of quasi-experiments in applied fields such as occupational therapy has led to criticism of the resulting research literature (Williamson, Goldschmidt, & Colton, 1986). Detractors frequently argue that the reported studies are methodologically weak and scientifically uninterpretable. These criticisms have contributed to a disenchantment with traditional experimental approaches in applied fields and have encouraged a search for alternative methodologies.

Explication and Experimentation

Numerous authors have examined the published literature in medicine, education, and the behavioral and social sciences and commented on the failure of the reported designs to meet the standards of the experimental model (DerSimonian, Charietle, McPeek, & Mosteller, 1982; Dunst & Rheingrover, 1981). These methodological reviews typically examine a series of studies either from a specific journal or addressing a particular research question. For example, Arendt, MacLean, and Baumeister (1988) assessed the methodological quality of eight studies evaluating the effectiveness of sensory integration treatment procedures with subjects with mental retardation. The authors examined such areas as methodological quality, subject assignment, methods of data collection, and types of data analysis. On the basis of the ratings, Arendt et al. concluded that the studies were poorly designed, were scientifically uninterpretable, and provided no experimental support for the therapeutic use of sensory integration activities with patients with mental retardation.

Methodological reviews in applied fields invariably reveal that the majority of published clinical research studies are methodologically or statistically inadequate. The most comprehensive methodological critique was reported by Williamson et al. (1986), who analyzed 28 methodological review articles examining the scientific adequacy of study design, statistical methods, and documentation in 4,200 biomedical research reports. The 28
methodological review articles included in their analysis used a variety of design criteria to establish the scientific quality of the research articles examined. Williamson et al. reported that “poor scientific quality was found in nearly all the research—regardless of type and content” (p. 382). For example, the 12 methodological review articles published during or after 1970 assessed the scientific quality of 2,172 publications. In analyzing these articles, Williamson et al. found that only 6% of the studies met the minimum design and statistical criteria established by the authors of the 12 methodological reviews. Williamson et al. summarized their findings by observing that clinical researchers and practitioners “will find relatively few journal articles that are scientifically sound in terms of reporting usable data or providing even moderate support for their inferences” (p. 303).

The implication in the Williamson et al. (1986) review and in similar analyses of the clinical research literature is that studies that do not meet some minimum standard of methodological (design) rigor are uninterpretable from a scientific viewpoint. It is important to note that Williamson and colleagues did not state that the interventions examined were ineffective, but rather, that the results of the studies with various design flaws, such as lack of random assignment, were uninterpretable from a scientific viewpoint.

Design quality factors, such as random assignment to groups and blind recording of the dependent variable, influence the interpretation of a single study. The use of such design factors in the interpretation of aggregated research results is more problematic. Research design has a logic of its own, but it is not a logic that is necessarily appropriate to the collective integration or interpretation of multiple studies. An isolated study that does not use random assignment, blind recording, or some other attribute of design quality may indeed produce results that are difficult to interpret. In this respect, Williamson et al. (1986) and other methodological critics are correct in their conclusion regarding interpretation. An isolated study with poor internal validity is essentially uninterpretable.

A collection of methodologically imperfect studies, however, can provide a convincing case for a treatment effect if one important consideration is met—that the same design (methodological) weakness does not exist in each of the studies. For example, suppose six studies are identical in terms of research question, overall procedures, and outcome. Three of the studies, however, use blind recording of the outcome measure, whereas the other three do not. If all six studies produce the same statistical result, what are we to conclude regarding the importance of blind recording as a measure of study quality in this series of research investigations? In theory, blind recording of the dependent variable is an excellent practice, but in this set of studies the design characteristic did not co-vary with the study outcome. It would be a mistake to discard or ignore the three studies without blind recording in evaluating the effectiveness of the independent variable.

Design attributes are too important to be conceptualized as predetermined categories that capture scientific quality. A focus on the absolute importance of selected design characteristics as essential to the interpretation of research results promotes a rigid approach in which research problems are subordinated to empirical methodology. Evidence of this attitude is apparent whenever a particular design or method is considered superior regardless of the research context or question.

The alternative argument is that the effect of particular design factors on study outcomes should be examined empirically and not be considered an a priori assumption based on methodological convention. The purpose of the present investigation was to examine the influence of one attribute of experimental design on study outcomes in the occupational therapy research literature. The design characteristic examined was type of subject assignment, specifically, whether studies that used random assignment of participants to conditions produced different outcomes than studies that did not use random assignment. Random assignment of subjects to conditions is a vital component of traditional experimental methodology (Cook & Campbell, 1979) and is often used as an indicator of the quality of a research design. In discussing the importance of random assignment to the quality of clinical research, Kramer and Shapiro (1984) stated that “randomization of individual subjects has become firmly enshrined as a sine qua non of a methodologically sound clinical trial” (p. 2740).

Method

To examine covariation between type of subject assignment and study outcome, I looked at 44 articles from the American Journal of Occupational Therapy (AJOT) and the Occupational Therapy Journal of Research (OTJR). All 44 articles involved a parallel-groups comparison design, also referred to as a pretest-posttest control group design (Campbell & Stanley, 1963). This design incorporates a between-subjects comparison in which performance on a dependent measure is statistically compared across treatment and control comparison groups. Twenty-two of the articles employed random assignment of subjects into groups or conditions, and 22 used a nonrandom process. Examples of nonrandom methods of subject assignment included some form of matching on specific participant characteristics or the use of pre-existing groups or volunteers.

Study Retrieval

The 44 articles included in the analysis were identified through a review of individual studies in AJOT and OTJR.
beginning with the last issue in the 1990 volume year and working backward through previous issues. Eleven articles using a pretest-posttest control group design with random assignment and 11 articles using a similar design with nonrandom assignment were identified in each journal. The range of issues reviewed was from 1981 through 1990 for OTJR and from 1980 through 1990 for AJOT.

The nature of the subjects' characteristics and the research question or independent variable were not considered in the identification and selection process. All articles retrieved, however, involved a comparative research question related to the effectiveness of some therapeutic intervention. For example, Poole, Whitney, Hangeland, and Baker (1990) recently reported the results of a true experimental design in which 18 subjects with hemiplegia were randomly assigned to splint or non splint treatment conditions. The subjects in the splint condition received 30 min of treatment with an inflatable splint as part of their occupational therapy program. An analysis of variance revealed no statistically significant differences between the treatment and control groups on measures of range of motion, pain, and motor function following a 3-week intervention period.

Study Coding

The 44 articles were coded by two raters with research experience and backgrounds in the statistical analysis and interpretation of clinical research investigations. The information coded from each article included the number and names of the authors, year of publication, source of publication, sample size, type of dependent variable, type of statistical value, reported probability values, method of subject assignment, and method of recording the dependent variable. Interrater reliability was computed for all coding categories with the Kappa statistic and produced values ranging from .89 to 1.00. (Copies of the coding frame and a listing of the 44 articles may be obtained from the author.)

Results

The median number of subjects per study was 28. Four studies had fewer than 15 subjects, and 2 had more than 100. A total of 381 statistical tests were contained in the 44 studies. The mean number of statistical hypotheses (8.66) compares closely with previous examinations of the biomedical and rehabilitation research literature (Ottenbacher, 1988; Ottenbacher & Barrett, 1990; Pocock, Hughes, & Lee, 1987). Sixty-nine (18.1%) of the statistical hypotheses tests conducted in the 44 investigations were reported as statistically significant. A contingency table analysis examining statistical outcome (significant vs. nonsignificant) by type of design (randomized vs. nonrandomized) resulted in a chi-square of 7.83 ($p < .01, df = 1$). This finding suggests that statistically significant results were more frequently associated with designs that did not use random assignment. That is, studies with a methodological limitation were more frequently associated with a positive statistical outcome.

This initial analysis suggests that poor designs (those without random assignment) appear to be affiliated with an increased likelihood of producing statistically significant results. There are, however, several serious limitations with this analysis and the subsequent implication that a design artifact (type of subject assignment) is associated with statistically positive results. First, the outcome of a statistical test (significance vs. nonsignificance) is strongly related to sample size. Sample size may in turn be related to type of research design. For example, quantitative reviews (meta-analyses) of applied research questions have frequently reported that studies rated high in design quality factors (e.g., use of random assignment) have smaller sample sizes than those rated as having weaker designs (Ottenbacher & Cooper, 1983; Ottenbacher & Peterson, 1995). In the present investigation, the mean number of subjects in the 22 randomized studies was 22.44 ($SD = 7.90$). The mean sample size for the nonrandomized studies was 32.40 ($SD = 11.71$). The difference between the mean sample sizes was statistically significant ($p < .05, df = 42$).

Second, the 381 statistical tests included in the reviewed studies do not represent independent data points. Multiple statistical tests were obtained from individual studies. For example, one nonrandomized study may have reported 10 positive statistical tests, whereas a randomized investigation may have been associated with a single statistical test that was nonsignificant. For an examination of the relationship between study design attributes (i.e., randomized vs. nonrandomized trials) and statistical outcome, a method is needed to quantify study findings that avoids the problem of dependence and reduces or eliminates the influence of sample size.

An unbiased estimate of study effect size, the $d$-index, was computed for each of the primary hypotheses under evaluation in the 44 studies. The $d$-index is used to estimate the difference between two groups in terms of their common (average) standard deviation. For example, if $d = .30$, then $3/10$ of a standard deviation separates the average subject in the two groups. The $d$-index transforms the results from any two-group comparison into a standardized metric.

Cohen (1988) presented a detailed description of various effect-size metrics as well as tables to compute power and sample size. The $d$-index was used in this illustrative analysis because it is simple to compute, it is scalefree, and it is the appropriate effect size for use with a two-group comparison generated from the parallel-groups comparison design (Cohen, 1988; Lipsey, 1990).

Effect sizes ($d$-indexes) can be compared from means and standard deviations for the experimental and
control or comparison groups or from \( t \) and \( F \) ratios when the means and standard deviations are not reported in the article. Several authors have provided formulas and rationales for transforming \( t \) and \( F \) values to d-indexes (Friedman, 1968; Kraemer & Thieman, 1987).

In studies where exact \( t \) and \( F \) values were not provided, they were estimated from significance levels and sample size. When nonparametric statistical techniques were reported, effect-size estimates were computed with procedures described by Hedges and Olkin (1985). In the present study, when an investigation reported a statistically nonsignificant result but not enough information to determine the d-index, an effect size of 0.00 was assumed. This is a conservative estimate indicating that there was no difference in performance across the groups being compared (Ottenbacher & Barrett, 1989).

Effect-size estimates (d-indexes) were computed for each of the primary hypotheses in the 44 studies. In those studies that involved more than one primary hypothesis, the average d-index was computed and reported. The mean effect size for the 22 AJOT articles was 0.38 \((SD = 0.18)\); for the 22 studies from the OTJR, 0.38 \((SD = 0.14)\). Because these values were not statistically different \((p > .05, df = 42)\), the d-indexes were collapsed and subsequent analyses were performed on all 44 effect-size values.

The mean effect sizes for the two design categories are shown in Table 1. Also included in Table 1 are the weighted mean effect sizes, the standard deviation, and the 95% confidence intervals for the weighted mean effect sizes. The weighted mean effect sizes are the product of each d-index weighted by the sample size of the study (Hedges & Olkin, 1985). Construction of the 95% confidence intervals revealed that the mean effect sizes across the two design categories did not differ significantly.

### Discussion

Numerous articles and tutorials have appeared in the rehabilitation and health-related literature providing guidelines for the interpretation and critique of research studies (Ethridge & McSweeney, 1971; Louis & Shapiro, 1983; Rieglman & Hirsh, 1989). These methodological checklists serve a purpose in educating readers regarding various components of the research process as they relate to the conducting and interpreting of a single study. They are relatively less informative, however, when the question turns from interpretation of a single study to evaluation of a series of studies addressing a specific area of research. The results of the present investigation suggest that judgments regarding how design quality variables (such as random assignment) interact with study outcome should be addressed empirically and not based on a priori assumptions derived from a predetermined design hierarchy (Cronbach, 1982).

#### Table 1

<table>
<thead>
<tr>
<th>Design</th>
<th>AJOT</th>
<th>OTJR</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random assignment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( N )</td>
<td>11</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>Mean effect size</td>
<td>.36</td>
<td>.39</td>
<td>.38</td>
</tr>
<tr>
<td>Mean weighted effect size</td>
<td>.34</td>
<td>.35</td>
<td>.35</td>
</tr>
<tr>
<td>( SD )</td>
<td>1.17</td>
<td>1.14</td>
<td></td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>.25-47</td>
<td>.32-46</td>
<td>.32-44</td>
</tr>
<tr>
<td>Nonrandom assignment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( N )</td>
<td>11</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>Mean effect size</td>
<td>.39</td>
<td>.37</td>
<td>.39</td>
</tr>
<tr>
<td>Mean weighted effect size</td>
<td>.36</td>
<td>.35</td>
<td>.36</td>
</tr>
<tr>
<td>( SD )</td>
<td>.19</td>
<td>.16</td>
<td>.18</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>.27-51</td>
<td>.27-47</td>
<td>.31-47</td>
</tr>
</tbody>
</table>

Note: \(AJOT = \) American Journal of Occupational Therapy. \( OTJR = \) Occupational Therapy Journal of Research.

#### Dogmatism and Design

A common assumption is that studies not meeting some predetermined design criterion will be biased in favor of the independent variable and, therefore, will produce spurious positive results (Williamson et al., 1986). This assumption has been operationalized by Sackett (1986) in a strategy to facilitate the interpretation of multiple studies. The model presented by Sackett includes a design hierarchy based on levels of evidence. The scheme provides five levels of evidence related to an investigation's design quality and ability to control experimental (Type I and Type II) errors. Three grades of recommendation are possible based on the level of evidence. Sackett's model provides a framework for the synthesis of studies that address similar hypotheses but that vary in design attributes. Investigations associated with Level I or II evidence are given more weight in the development of clinical recommendations than are studies associated with Level III, IV, or V evidence.

In his levels of evidence model, Sackett (1986) assumed that studies with flaws in design quality, that is, Levels III, IV, and V, will provide biased results that favor the independent variables; studies of poor design quality, therefore, are more likely to produce positive results. Specifically, Sackett stated, "Treatments are much less likely to be judged efficacious in double blind, randomized trials than in uncontrolled case studies, or unblinded 'open' comparisons with contemporaneous controls" (p. 29).

#### Implications

The results of the present investigation suggest that this assumption is not universally warranted, particularly when measures of study outcome that are not sample-size sensitive are used (Colditz, Miller, & Mosteller, 1989). In fact, the opposite assumption regarding study outcome could be made. That is, studies with poor designs...
studies with strong designs will be more sensitive to the statistical results of the evaluation in sufficient detail examining the effectiveness of psychotherapy interven­

tion. Weiss and Weisz (1990) recently provided an excellent example of this phenomenon in their analysis of the results of 105 studies examining the effectiveness of psychotherapy intervention for children. All studies included a comparison between a treatment and a control group and presented the statistical results of the evaluation in sufficient detail to compute effect-size measures. Eight methodological variables were coded for the 105 studies. The authors used a general linear model to examine the influence of various quality and validity factors on study outcome. The analysis revealed that those studies with the strongest designs (as measured by the highest validity scores) were most often associated with positive statistical results. They concluded their investigation by stating, “In general, increased experimental rigor was related to larger effect sizes; this argues against the hypothesis that methodologically weak studies have led to an overestimate of therapy effects” (Weiss & Weisz, 1990, p. 639).

The influence of design quality factors should not be an a priori assumption, but rather, an empirical question to be systematically explored whenever more than two studies are synthesized. If there is a covariation between a particular design attribute and study outcome, then this characteristic must be considered in the evaluation and interpretation of the aggregated study results. For example, if studies that use random assignment to conditions are associated with results that are more or less favorable than studies that do not include this design characteristic, then the design attribute must be considered in any attempt to evaluate and synthesize related studies. If, however, no relationship exists between a particular design characteristic and study outcome, then studies with and without the design feature should not be treated differently within the context of the research literature being examined. We can make this empirical judgment by considering research design as a moderator variable and examining its contribution to aggregated research outcomes as we would any other important moderator variable (cf. Ottenbacher, 1991, in press-a, in press-b).

**Conclusion**

The result of this investigation should not be interpreted as an argument against the consideration of design quality factors in the interpretation and synthesis of clinical research: Design attributes are of the utmost importance. In fact, design considerations are too important to be conceptualized as predetermined categories that capture empirical quality. An attitude that views one design or research approach as the best, regardless of the context or question, reflects an empirical parochialism that restricts legitimate scientific inquiry. It also creates a disillusionment with conventional research procedures, a disillusionment that is not methodologically warranted and that may result in a misinterpretation of research findings and procedures. Unfortunately, this sense of disillusionment has led many occupational therapists to view experimental design and statistical analysis as partners in a charmed circle of reductionistic inquiry that has little relevance beyond its own circumference.

One positive outcome of the disillusionment with conventional research methods has been the recognition and identification of alternative research approaches. The search for alternative ways of knowing is an important and useful aspect of the productive chaos referred to earlier (Cook & Campbell, 1979).

The fact that conventional research procedures have been found to contain certain limitations does not mean that they should be rejected as irrelevant to the development of knowledge in applied fields such as occupational therapy. Abandonment of approaches that are not appropriately implemented or properly interpreted may result in replacement of experimental dogmatism with phenomenological orthodoxy (Hasselkus, 1991).

Carlson and Clark (1991) recently presented an excellent overview and discussion of evolving epistemology in occupational therapy. If we follow their advice, we can remove some of the rigidity and ritual from our research endeavors and begin to examine the logic of our empirical actions and inferences. The rigidity associated with experimental methods has led some researchers to use design and statistics like a drunken man uses a lamp-post—more for support than for illumination. By recognizing the dynamic nature of design quality and relying less on stereotyped statistical inferences, we may lose some support but, in the long run, will see our research findings more clearly as a result of the improved empirical illumination.

---

**References**


