Exceptional Children

Vol. 71, No.2, pp. 181-194.

Evaluating the Quality of Evidence From Correlational Research for Evidence-Based Practice

BRUCE THOMPSON
Texas A&M University and Baylor College of Medicine

KAREN E. DIAMOND
Purdue University

ROBIN MCWILLIAM
Vanderbilt University

PATRICIA SNYDER
Louisiana State University Health Science Center

SCOTT W. SNYDER
University of Alabama

ABSTRACT: *Only true experiments offer definitive evidence for causal inferences, but not all educational interventions are readily amenable to experiments. Correlational evidence can at least tentatively inform evidence-based practice when sophisticated causal modeling or exclusion methods are employed. Correlational evidence is most informative when exemplary practices are followed as regards (a) measurement, (b) quantifying effects, (c) avoiding common analysis errors, and (d) using confidence intervals to portray the range of possible effects and the precisions of the effect estimates.*

In their recent article in the *Educational Researcher*, Feuer, Towne, and Shavelson (2002) asked,

> What are the most effective means of stimulating more and better scientific educational research?... [T]he *primary emphasis* [italics added] should be on nurturing and reinforcing a scientific culture of educational research. (p. 4)

They defined scientific culture as "a set of norms and practices and an ethos of honesty, openness, and continuous reflection, *including how research quality is judged*" (Feuer, Towne, & Shavelson, 2002, p. 4, italics added). Recent movements to emphasize evidence-based practice in medicine (see Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000), psychology (see Chambless, 1998), and education (Mosteller & Boruch, 2002; Shavelson & Towne, 2002) also reflect the

necessity for standards with which to evaluate research evidence, including evidence from correlational designs.

WHAT IS CORRELATIONAL EVIDENCE?

Correlational studies can be defined in various ways. In one sense, *all* analyses are correlational (Cohen, 1968; Knapp, 1978; Thompson, 2000a). Because all conventional parametric analyses (e.g., *t*-tests, ANOVA, ANCOVA) are correlational (Bagozzi, Fornell, & Larcker, 1981), in a sense every quantitative study yields correlational evidence. Distinguishing evidence types by focusing on the analysis is not useful because under such a broad definition, all evidence would fall under this single umbrella. Furthermore, a given analysis (e.g., multiple regression) can be correctly employed to analyze data from numerous designs (e.g., a true experiment, a comparative design). A more useful distinction regarding types of evidence focuses not on the analysis, but on the *design* of the study yielding the evidence.

Correlational studies are *quantitative, multisubject designs in which participants have not been randomly assigned to treatment conditions.* Analytic methods commonly (but not exclusively) applied with such designs are multiple regression analysis, canonical correlation analysis, hierarchical linear modeling, and structural equation modeling.

For example, as defined here, a correlational study might investigate differential achievement levels of students enrolled in classes of different sizes, where the students were not randomly assigned to classes of given sizes. Or researchers might collect data regarding the frequency with which teachers praise students, to examine relationships of these behaviors with students' self-concepts and school attendance.

HOW CAN CORRELATIONAL EVIDENCE INFORM PRACTICE?

Definitive causal conclusions in quantitative research can only be reached on the basis of true randomized trials. That is why it is so important for educational researchers to conduct more true experiments. Historically, randomization has been too infrequently invoked within the social sciences (Ludbrook & Dudley, 1998). However, for various reasons, evidence from types of research not involving randomized clinical trials is also relevant to evidence-based practice.

It is crucial to match research questions and research designs, and some questions are best addressed with nonexperimental designs. For example, questions involving school or classroom culture may require qualitative methods, and questions involving the intensive study of learning dynamics of individual children may require single-subject studies. Even when group quantitative methods are appropriate, randomized experiments may not be ideal if the immature state of knowledge on a given issue does not yet justify the expense of such trials. And in some cases clinical trials may raise ethical questions regarding denial of needed services to control group participants. Not all questions can be addressed with clinical trials, and unduly widespread use of clinical trials would also be undesirable because cross-contamination of effects across children involved in multiple experiments would then compromise all results.

Correlational designs do not provide the best evidence regarding causal mechanisms. Nevertheless, in at least two ways correlational evidence can be used to inform causal inferences and thus evidence-based practice. The first approach is *statistically based*, and involves statistically testing rival alternative causal models, even though the design is correlational. The second method is *logic based* and invokes logic and theory with nonexperimental data in an attempt to rule out all reasonable alternative explanations in support of making a single plausible causal inference.

Statistical Testing of Rival Causal Models

The analytic methods that today we call structural equation modeling (SEM; or covariance structure analysis) originated in the work of Karl Jöreskog (e.g., 1969, 1970, 1971, 1978), and the computer program, LISREL (i.e., analysis of *LI*near *S*tructural *REL*ationships) developed by Jöreskog and his colleagues (e.g., Jöreskog & Sörbom, 1989). These methods as originated in the 1960s and 1970s were then often called *causal modeling*,

which hints at the potential of SEM to inform causal inferences.

SEM incorporates factor (or measurement) models, building on the factor analytic methods proposed by Spearman (1904), and a structural model linking these latent constructs, building on the path analytic methods proposed by Wright (1921, 1934). Within the structural model, analysts may test whether (a) two latent constructs (X and Y) covary or are correlated, (b) X causes Y, (c) Y causes X, or (d) X and Y reciprocally cause each other.

The appeal of SEM is that rival models can, and indeed should, be tested (see Thompson, 2000b). If only one of these four models fits the data (e.g., a model specifying that X causes Y), then there is at least some evidence bearing on the existence of a causal relationship.

For example, data reported by Bagozzi (1980) have been used in several reports to illustrate this application (see Jöreskog & Sörbom, 1989, pp. 151–156, and Thompson, 1998b, pp. 37–39). Bagozzi's study investigated the job satisfaction and job performance of 122 workers. For these data, it appeared that a model positing that job performance leads to job satisfaction better fit the data than did models positing that job satisfaction leads to job performance or that satisfaction and performance are reciprocally related.

Logically-Based Exclusion Methods

In some cases when true experiments are not performed, and even when structural modeling is not used, we still may be able to reach causal inferences with some degree of confidence. The capacity for extracting causal information from nonexperimental designs (e.g., intervention studies not invoking random assignment to groups) turns on our capacity to evaluate whether all relevant preintervention differences and design validity threats can be excluded (i.e., deemed essentially irrelevant).

For example, let's say two intact (i.e., not randomly assigned) groups of special education students were taught reading with two different curricula. We want to make some causal interpretation of the postintervention reading differences in the two conditions. We might investigate preintervention differences in the students on everything that we consider as being even potentially relevant (e.g., preintervention reading scores, socioeconomic status). We might also try to confirm that there were no meaningful extraneous contaminants of treatment influences (e.g., teachers had similar backgrounds in both conditions, curricula were implemented with fidelity). If we can rule out all such problems, we may have at least some plausible evidence that one curriculum is superior to the other curriculum, even though we have not performed a true experiment, and we have not statistically tested rival causal models.

Definitive causal conclusions in quantitative research can only be reached on the basis of true randomized trials.

The challenge to such efforts is that we may not be certain exactly which preintervention differences or what design validity threats are relevant in a given study. The beauty of true experiments is that the law of large numbers creates preintervention group equivalencies on *all* variables, even variables that we do not realize are essential to control.

But exclusion methods may be necessary in an environment where true experiments can not be used to address every important intervention question. And as our knowledge base grows, we may become more certain regarding which preintervention differences or treatment confounds are most noteworthy.

Limitations of Nonexperimental Research

Both statistical modeling and logical exclusion methods require that models are "correctly specified." That is, the analytic results are sound only to the extent that

- All the correct variables, and only the correct variables, are employed within the tested models.
- The correct dynamics (e.g., mediation, moderation) are specified within the tested models (i.e., the correct analysis is used).

But as Pedhazur (1982) has noted, "The rub, however, is that the true model is seldom, if ever, known" (p. 229). And as Duncan (1975) has

noted, "Indeed it would require no elaborate sophistry to show that we will never have the 'right' model in any absolute sense" (p. 101). Thus, both methods must be used cautiously in applying correlational evidence to help inform evidence-based practice. Nevertheless, correlation evidence, like other nonexperimental evidence, is relevant to evidence-based practice.

PURPOSE OF THE PRESENT ARTICLE

Educational research has sometimes been criticized for being poorly conducted (see Gall, Borg, & Gall, 1996, p. 151). For example, the National Academy of Science evaluated educational research generically and found "methodologically weak research, trivial studies, an infatuation with jargon, and a tendency toward fads with a consequent fragmentation of effort" (Atkinson & Jackson, 1992, p. 20). Nevertheless, even imperfect studies may provide some useful information. Few defects in published studies are sufficiently egregious to warrant total disqualification from any consideration.

A possible exception to this generalization encompasses studies using stepwise methods (Snyder, 1991). As Huberty (1994) noted, "It is quite common to find the use of 'stepwise analyses' reported in empirically based journal articles" (p. 261). Thompson (1995, 2001) explained that stepwise methods (a) do not correctly identify the best subset of predictors, (b) yield results that tend to be nonreplicable, and (c) "are positively satanic in their temptations toward Type I errors" (Cliff, 1987, p. 185), because most computer programs incorrectly compute the degrees of freedom for stepwise analyses. When researchers must select a subset of variables from a larger constellation of choices, the "all-possible-subsets" analyses described by Huberty (1994), and available in SAS, provide reasonable results.

The present article proposes some quality indicators for evaluating correlational research in efforts to inform evidence-based practice. Given the inherent challenges of educational research (Berliner, 2002), most studies are unavoidably imperfect and vary in the quality of the evidence they provide.

The quality indicators proposed are not new. But they may be insufficiently honored in contemporary analytic practice. For example, various effect size statistics have been proposed for decades (Huberty, 2002), but studies have shown effect sizes to be reported in less than half the published articles in various journals and various disciplines (see Thompson, 1999b; Vacha-Haase, Nilsson, Reetz, Lance, &, Thompson, 2000). Similarly, confidence intervals have been recommended for years (see Chandler, 1957), but empirical studies suggest that intervals are infrequently reported in published social science research (Kieffer, Reese, & Thompson, 2001).

The quality indicators presented are grouped into four sets: (a) measurement; (b) practical and clinical significance; (c) avoidance of some common analytic mistakes; and (d) confidence intervals for score reliability coefficients, statistics, and effect sizes. These are not the only indicator categories that might be identified, but the present categories will serve reasonably well to distinguish some recognizable features of correlational inquiry. Where space limitations preclude in-depth exploration of concerns, helpful references providing further elaboration are routinely provided.

MEASUREMENT

The quality of the evidence informing practice is inherently limited by the psychometric integrity of the data being analyzed in a given study. Classically, measurement concerns are conceptualized as involving two primary considerations: score reliability and score validity. However, some modern measurement theories actually present a unified view of these concerns, such that reliability and validity issues are blended (Brennan, 2001).

Reliability can be conceptualized as addressing the question, "Do the scores measure anything?" (i.e., are nonrandom), and *validity* addresses the question, "Do the scores measure only the correct something that they are supposed to measure?" (Thompson, 2003). In this classical measurement view, reliability is a necessary but insufficient condition for validity.

Researchers have traditionally recognized that score validity is not immutable within a

given measure; the same measure may yield scores valid for some purposes and respondents, and invalid for other inferences or respondents (Schmidt & Hunter, 1977). Lately, more researchers have come to realize in a similar vein that a given test also is not immutably reliable. As Wilkinson and the American Psychological Association (APA) Task Force on Statistical Inference (1999) recently emphasized:

> It is important to remember that a test is not reliable or unreliable.... Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. (p. 596)

Unfortunately, recent empirical studies of published research reports indicate that the vast preponderance of articles do not even mention reliability, much less report reliability for the data actually being analyzed (Vacha-Haase, Henson, & Caruso, 2002). These practices may originate in misconceptions that tests are reliable, and that once reliability has been established in a given sample, further concerns are moot (Thompson, 2003; Vacha-Haase, 1998).

A problematic practice is to "induct" the reliability coefficient from a prior study or a test manual (Vacha-Haase et al., 2002). Unfortunately, this induction of prior reliability coefficients turns on the premises that (a) the samples are comparable in their compositions and (b) the scores are roughly equivalent in their standard deviations across studies (Crocker & Algina, 1986, p. 144). Sadly, empirical studies suggest that such inductions are almost never explicitly justified and often are wildly inappropriate (Vacha-Haase, Kogan, &, Thompson, 2000; Whittington, 1998). It is unacceptable to induct the score reliability coefficients from prior studies or test manuals if there is no *explicit* evidence presented that the sample compositions and the standard deviations from the prior study and a current study are *both* reasonably comparable.

Quality Indicators:

- Score reliability coefficients are reported for all measured variables, based on induction from a prior study or test manual, with explicit and reasonable justifications as regards comparabilities of (a) sample compositions and (b) score dispersions.

Recent empirical studies of published research reports indicate that the vast preponderance of articles do not even mention reliability, much less report reliability for the data actually being analyzed.

- Score reliability coefficients are reported for all measured variables based on analysis of the data in hand in the particular study.
- Evidence is inducted, with explicit rationale, from a prior study or test manual that suggests scores are valid for the inferences being made in the study.
- Score validity is empirically evaluated based on data generated within the study.
- The influences of score reliability and validity on study interpretations are explicitly considered in reasonable detail.

PRACTICAL AND CLINICAL SIGNIFICANCE

Statistical significance estimates the probability, *p*, of sample results, given the sample size, and assuming the sample came from a population exactly described by the null hypothesis (Cohen, 1994; Thompson, 1996). In disciplines as diverse as wildlife sciences and psychology, the utility of statistical significance has been increasingly questioned in recent years (Anderson, Burnham, & Thompson, 2000; see Harlow, Mulaik, & Steiger, 1997 and Nickerson, 2000 for comprehensive summaries of both sides of the controversy). Indeed, a forthcoming issue of the *Journal of Socio-Economics* (see Thompson, in press-b) will include commentary by several economics Nobel laureates on this issue.

Practical significance evaluates the potential noteworthiness of study results, by quantifying the degree to which sample results diverge from the null hypothesis (Snyder & Lawson, 1993). These quantifications are often referred to generically as "effect sizes." There are literally dozens of effect size statistics (see Kirk, 1996). Many of these myriad choices can be arrayed within the

following categories: (a) standardized differences (e.g., Cohen's *d*, Glass's Δ), (b) "uncorrected" variance-accounted-for (e.g., η^2, R^2), and (c) "corrected" variance-accounted-for (e.g., adjusted R^2, ω^2; see Thompson, 2002a). Thompson (in press-a), Kirk (1996), and Snyder and Lawson (1993) provide reviews on the numerous available choices.

Clinical significance evaluates the extent to which intervention recipients no longer meet diagnostic criteria (e.g., learning disability, depression), and, thus, no longer require specialized intervention (Jacobson, Roberts, Berns, & McGlinchey, 1999; Kendall, 1999). Clinical significance is potentially relevant *only* when the outcome variable can be interpreted using accepted diagnostic criteria (e.g., total cholesterol greater than 200 milligrams).

The fifth edition of the APA (2001) *Publication Manual* emphasized that

> It is *almost always necessary* to include some index of effect size or strength of relationship.... The general principle to be followed...is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. (pp. 25–26, emphasis added)

The manual also describes failure to report effect sizes as "a defect" (p. 5). But the editors of 23 journals have gone beyond the APA *Publication Manual* and have published author guidelines requiring effect size reporting (Fidler, 2002).

Jacob Cohen, in his various books on power analysis, provided benchmarks for effect sizes that he deemed small, medium, and large. He formulated these based on his impressions of the range of effect sizes typical of the social science literature as a whole. He hesitated to provide such benchmarks because he felt that effects ought to be interpreted instead against the criteria of the researcher's values and related effects reported in prior literature. However, he provided these benchmarks of typicality because he felt that researchers would be more likely to report effect sizes if there were some standards for interpreting them, pending the reporting of effect sizes becoming routine within the literature. But "if people interpreted effect sizes [using fixed benchmarks] with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (Thompson, 2001, pp. 82–83).

Glass, McGaw, and Smith (1981) argued that "there is no wisdom whatsoever in attempting to associate regions of the effect-size metric with descriptive adjectives such as 'small,' 'moderate,' 'large,' and the like" (p. 104). The only exception to this rule involves groundbreaking inquiry in which little or no previous research has been conducted, in which case Cohen's benchmarks may be useful as a (very) rough guide.

The problem is *not* the use of adjectives such as large or small. The problem is using fixed, generic benchmarks for making these judgments, rather than consulting the effects in related studies.

The results of a single study have meaning primarily as regards what they contribute to a literature, although, of course, the results of a single study sometimes do change thinking about a phenomenon (Thompson, in press-a). The comparison of effects against those reported in related prior studies enables researchers to evaluate the consistency of results across studies. This powerful view of all quantitative research as requiring "meta-analytic thinking" (Cumming & Finch, 2001; Thompson, 2002b) is promoted by interpreting results across studies. Such direct comparisons also alert researchers to inconsistent findings, which may highlight moderator variables or situations in which results vary across different subpopulations.

Common Mistakes

Even today when 23 journals (see McLean & Kaufman, 2000; Snyder, 2000) require effect size reporting, effect size reporting is more the exception than the norm (see Thompson, 1999b; Vacha-Haase, Nilsson et al., 2000). This does make it more difficult to interpret the effects in a given study in direct, explicit comparison with the effect sizes reported in prior studies, because effects must be computed or estimated for prior studies in which authors did not report effect sizes.

Effect sizes should be reported for all primary study outcomes, even when particular results are not statistically significant (Thompson,

2002b). Such reporting facilitates future meta-analytic integration of the study into the corpus of the literature.

Furthermore, some researchers do report, but do not interpret, their effect sizes (Vacha-Haase, Nilsson et al., 2000). Reporting, but not interpreting, effect sizes does not allow effect sizes to inform fully the interpretation of results.

A fundamental, but too common, mistake is failing to identify which effect size is being reported. Because there are so many different effect sizes (see Kirk, 1996), some with different ranges and properties, it is critical to identify reported effect statistics explicitly.

Finally, it is also important to recognize that effect sizes cannot magically escape the limitations or analytic assumptions of given analyses (Olejnik & Algina, 2000). These limitations and assumptions should be considered as part of result interpretation.

Quality Indicators:

- One or more effect size statistics is reported for each study primary outcome, and the effect statistic used is clearly identified.
- Authors interpret study effect sizes for selected practices by directly and explicitly comparing study effects with those reported in related prior studies.
- Authors explicitly consider study design and effect size statistic limitations as part of effect interpretation.

AVOIDANCE OF SOME COMMON MACRO-ANALYTIC MISTAKES

Across the literature a range of analytic errors are seen with some frequency. Some of these errors are unique to a particular method. For example, it is common for researchers to confuse descriptive discriminant analysis with predictive discriminant analysis, or vice versa, and consequently to misinterpret their discriminant analysis results (see Huberty, 1994; Kieffer et al., 2001). Other analytic errors occur generally across analytic choices.

Research evidence better informs practice when these errors are avoided. Here four such common generic, macro-analytic errors are noted. These can occur across two or more, or in some cases, all correlational analytic methods.

Failure to Interpret Structure Coefficients

Throughout the general linear model (GLM), weights are either explicitly (e.g., regression, descriptive discriminant analysis) or implicitly (e.g., *t*-tests, ANOVA) applied to measured variables to estimate the scores on the latent variables that are actually the focus of the analysis (see Thompson, 2000a). These weights are given different names across analyses (e.g., beta weights, factor pattern coefficients, discriminant function coefficients), which has the effect of obfuscating the existence of the GLM.

When researchers obtain noteworthy effects, they commonly (and correctly) consult these weights as part of the process of determining the origins of detected effects. However, these weights are usually *not* correlation coefficients of predictors with outcome variables. In fact, a predictor may have the largest nonzero weight in an analysis even when the predictor is perfectly uncorrelated with the outcome variable (Thompson & Borrello, 1985).

Structure coefficients (i.e., correlations of measured variables with the latent variables actually being analyzed, such as regression $\hat{Y}$ scores) are also usually *essential* to correct interpretation (Courville & Thompson, 2001; Dunlap & Landis, 1998). For example, structure coefficients have been characterized as essential to the correct interpretation of multiple regression analysis (Courville & Thompson), exploratory factor analysis (Gorsuch, 1983, p. 207), confirmatory factor analysis (Graham, Guthrie, &, Thompson, 2003), descriptive discriminant analysis (Huberty, 1994), and canonical correlation analysis (Thompson, 1984).

Quality Indicators:

- GLM weights (e.g., beta weights) are interpreted as reflecting correlations of predictors with outcome variables only in the exceptional case that the weights indeed are correlation coefficients.
- When noteworthy results are detected, and the origins of these effects are investigated, the in-

Effect sizes should be reported for all primary study outcomes, even when particular results are not statistically significant.

terpretation includes examination of structure coefficients.

Converting Intervally Scaled Variables to Nominal Scale

It is not uncommon (Pedhazur, 1982, pp. 452–453) to see researchers convert one or more of their independent or predictor variables into nominal scale in order to run OVA methods (e.g., ANOVA). For example, researchers may take intervally-scaled pretest data (e.g., IQ scores, pretest achievement scores) and characterize participants as either "low" or "high" in learning aptitude.

Such dichotomization (trichotomization, etc.) (a) "throws information away" (i.e., discards score variability; Kerlinger, 1986, p. 558); (b) attenuates reliability of the scores being analyzed; (c) distorts variable distributions; and (d) distorts relationships among variables (Thompson, 1986). The result is analyses that are ecologically less valid.

In her comprehensive Monte Carlo study, Hester (2000) provided considerably more detail on the consequences of such ill-considered analytic choices. The consequences of these conversions are particularly deleterious for building an integrated literature when different researchers use divergent cutoffs (e.g., different sample-specific median splits) to implement the conversions. For example, if researcher Jones dichotomizes pretest IQ data at Jones's sample median of 95, and researcher Smith does so at Smith's sample median of 105, we will never know whether discrepant ANOVA or MANOVA results are (a) an artifact of using different cutpoints to dichotomize, or (b) a failure to replicate results.

Quality Indicator:

- Interval data are not converted to nominal scale, unless such choices are justified on the extraordinary basis of distribution shapes, and the consequences of the conversion are thoughtfully considered as part of result interpretation.

Inappropriate Univariate Methods

Univariate methods (i.e., analyses using a single dependent variable) are quite commonly used in educational research (Kieffer et al., 2001). These methods can be quite appropriate for some studies. However, there are two situations in which univariate methods are inappropriate.

First, *univariate methods are generally inappropriate in the presence of multiple outcomes variables.* The use of univariate methods when a study involves several outcome variables (a) inflates the probability of experimentwise Type I errors, and (b) does not honor the reality that outcome variables can interact with each other to define unique outcomes that are more than their constituent parts (Fish, 1988).

Regarding the second concern, Thompson (1999a) provided a heuristic data set illustrating the importance of these issues. In his example data set, the two means on X and Y did not differ to a statistically significant degree (both ANOVA p values were .774), and furthermore the ANOVA eta^2 effect sizes were both computed to be 0.469%. Thus, the two sets of ANOVA results were not statistically significant, and they both involved extremely small effect sizes. However, a MANOVA/descriptive discriminant analysis (DDA) of the *same data* yielded a $p_{CALCULATED}$ value of .0002, and a multivariate eta^2 of 62.5%!

This means that the Bonferroni correction in the presence of several or many outcome variables is *not* suitable, for two reasons. First, the correction lowers power against Type II error. Second, multiple univariate analyses do not honor the ecological reality that all the variables, including the outcomes, can interact with each other to create unique effects that will only be discovered in a multivariate analysis.

Second, *the use of univariate methods (e.g., ANOVA) post hoc to multivariate tests is inappropriate,* albeit common (Kieffer et al., 2001). Put simply, a MANOVA and several ANOVAs each using the same measured outcome variables test completely different and irreconcilable effects, because the ANOVAs do not consider the relationships among the outcomes. These relationships are an essential consideration in the multivariate analyses, as illustrated in the Thompson (1999a) heuristic example.

In the words of Borgen and Seling (1978), "When data truly are multivariate, as implied by the application of MANOVA, a multivariate follow-up technique seems necessary to 'discover' the complexity of the data" (p. 696). It is illogical to first declare interest in a multivariate omnibus system of variables, and then to explore detected effects in this multivariate world by conducting nonmultivariate tests.

A logical MANOVA post hoc method is descriptive discriminant analysis, which Huberty (1994) noted is "closely aligned to the study of effects determined by a multivariate analysis of variance (MANOVA)" (p. 30). Huberty (1994) provided several chapters on using DDA post hoc to MANOVA to assess and describe multivariate dynamics.

Quality Indicators:

- Univariate methods are not used in the presence of multiple outcome variables.
- Univariate methods are not used post hoc to multivariate tests.

Failure to Test Statistical Assumptions

All statistical methods require that certain assumptions (e.g., homogeneity of variance in ANOVA, homogeneity of regression slopes in ANCOVA) must be met in order for *p* values and effect sizes to be accurate. Methodological assumptions are never perfectly met, but must be at least approximately met in order for results to be approximately correct.

Empirical studies of published articles suggest that statistical assumptions are too rarely tested by researchers (Keselman et al., 1998). These assumptions are more important than many researchers may realize, as suggested by Wilcox (1998) in his article titled, "How many discoveries have been lost by ignoring modern statistical methods?"

Statistical assumptions can be particularly important when statistical corrections are invoked, as in ANCOVA, particularly when used with nonrandom intact intervention groups (see Thompson, 1992). Using ANCOVA when the homogeneity of regression assumption is not met leads to "tragically misleading analyses" that actually "can mistakenly make compensatory education look harmful" (Campbell & Erlebacher, 1975, p. 597).

Quality Indicator:

- Persuasive evidence is explicitly presented that the assumptions of statistical methods are sufficiently well-met for results to be deemed credible.

CONFIDENCE INTERVALS FOR RELIABILITY COEFFICIENTS, STATISTICS, AND EFFECT SIZES

Confidence intervals (CIs) can be used to determine whether a given null hypothesis would be rejected. If a hypothesized value (e.g., $r = 0$; $r = .5$) is not within the interval, the null hypothesis positing the parameter value is rejected. However, this use of confidence intervals does not tap the primary positive features of using confidence intervals (Thompson, 1998a, 2001).

Confidence intervals inform judgment regarding all the values of the parameter that appear to be plausible, given the data (Cumming & Finch, 2001). Thus, by comparing the overlaps of confidence intervals across studies, researchers can evaluate the *consistency* of evidence across studies (Thompson, 2002b).

The widths of confidence intervals within a study, or across studies, also provide critical information regarding the *precision* of estimates in a study or in a literature. When intervals are wide, the evidence for a given point estimate being correct is called into question. Researchers may overinterpret effects in a literature and not recognize the imprecision of a body of literature, unless confidence intervals are computed and directly compared across studies.

For these various reasons, confidence intervals are increasingly recognized as being "in general, *the best* reporting strategy" and "the use of confidence intervals is therefore *strongly recommended*" (American Psychological Association, 2001, p. 22, emphasis added). Of course, as Fidler, Thomason, Cumming, Finch, and Leeman (2004) pointed out, as with any other statistical methods, CIs are not a panacea, and can be used thoughtlessly.

Univariate methods are generally inappropriate in the presence of multiple outcomes variables.

COMMON MISTAKES

Some researchers misinterpret confidence intervals as telling us how confident we may be (e.g., 95%) that a given, single interval captures a population parameter, such as a correlational effect size (e.g., r, r^2). However, the confidence statements when dealing with confidence intervals are about a large or infinite set of intervals drawn from a population capturing the interval a given percentage (e.g., 95%) of the time, and these confidence statements are *not* about single intervals (Thompson, 2002b).

We never know, unless we have the population data (and then would not be computing a CI), whether our single interval does or does not capture a population parameter. The probabilities of intervals capturing the population parameter (e.g., r, r^2) may be different even for a series of 95% confidence intervals.

Confidence intervals can be computed for (a) reliability coefficients (Fan & Thompson, 2001); (b) sample statistics (e.g., M, r); and (c) effect sizes (Thompson, 2002b). CIs are so appealing because using intervals across studies will ultimately lead us to the correct population value, even if our initial expectations are wildly in error (Schmidt, 1996)! Software for computing confidence intervals for effect sizes is widely available (Algina & Keselman, 2003; Cumming & Finch, 2001; Smithson, 2001; Steiger & Fouladi, 1992). Kline's (2004) recent book provides a comprehensive tutorial.

Quality Indicators:

- Confidence intervals are reported for the reliability coefficients derived for study data.
- Confidence intervals are reported for the sample statistics (e.g., means, correlation coefficients) of primary interest in the study.
- Confidence intervals are reported for study effect sizes.
- Confidence intervals are interpreted by direct and explicit comparison with related CIs from prior studies.

SUMMARY

Within the quantitative group-design genre, only true experiments offer definitive evidence for causal inferences that can inform evidence-based instructional practice. But not all educational interventions are readily amenable to experiments. In addition, experimental studies of educational interventions are compromised by cross-contamination when students participate in multiple interventions.

In such cases correlational evidence may be useful in adducing complementary evidence. Correlational studies can produce intriguing results that are then subjected to experimental study. And correlational evidence can at least tentatively inform evidence-based practice when sophisticated causal modeling (e.g., regression discontinuity analyses) or exclusion methods are employed. Correlational evidence is most informative when exemplary practices are followed with regard to (a) measurement, (b) quantifying effects, (c) avoidance of common macro-analytic errors, and (d) use of confidence intervals to portray the *consistency* of possible effects and the *precisions* of the effect estimates. Table 1 presents a list of the quality indicators suggested for research in this genre.

REFERENCES

Algina, J., & Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement, 63*, 537–553.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Anderson, D. R., Burnham, K. P., & Thompson, W. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*, 912–923.

Atkinson, R. C., & Jackson, G. B. (Eds.). (1992). *Research and education reform: Roles for the Office of Educational Research and Improvement*. Washington, DC: National Academy of Sciences. (ERIC Document Reproduction Service No. ED 343 961)

TABLE 1
Suggested Quality Indicators for Correlational Research

Measurement

1. Score reliability coefficients are reported for all measured variables, based on induction from a prior study or test manual, with explicit and reasonable justifications as regards comparabilities of (a) sample compositions and (b) score dispersions.
2. Score reliability coefficients are reported for all measured variables, based on analysis of the data in hand in the particular study.
3. Evidence is inducted, with explicit rationale, from a prior study or test manual that suggests scores are valid for the inferences being made in the study.
4. Score validity is empirically evaluated based on data generated within the study.
5. The influences of score reliability and validity on study interpretations are explicitly considered in reasonable detail.

Practical and Clinical Significance

6. One or more effect size statistics is reported for each study primary outcome, and the effect statistic used is clearly identified.
7. Authors interpret study effect sizes for selected practices by directly and explicitly comparing study effects with those reported in related prior studies.
8. Authors explicitly consider study design and effect size statistic limitations as part of effect interpretation.

Avoiding Some Common Macro-Analytic Mistakes

9. GLM weights (e.g., beta weights) are interpreted as reflecting correlations of predictors with outcome variables only in the exceptional case that the weights indeed are correlation coefficients.
10. When noteworthy results are detected, and the origins of these effects are investigated, the interpretation includes examination of structure coefficients.
11. Interval data are not converted to nominal scale, unless such choices are justified on the extraordinary basis of distribution shapes, and the consequences of the conversion are thoughtfully considered as part of result interpretation.
12. Univariate methods are not used in the presence of multiple outcome variables.
13. Univariate methods are not used post hoc to multivariate tests.
14. Persuasive evidence is explicitly presented that the assumptions of statistical methods are sufficiently well-met for results to be deemed credible.

CIs for Reliability Coefficients, Statistics, and Effect Sizes

15. Confidence intervals are reported for the reliability coefficients derived for study data.
16. Confidence intervals are reported for the sample statistics (e.g., means, correlation coefficients) of primary interest in the study.
17. Confidence intervals are reported for study effect sizes.
18. Confidence intervals are interpreted by direct and explicit comparison with related CIs from prior studies.

Bagozzi, R. P. (1980). Performance and satisfaction in an industrial sales force: An examination of their antecedents and simultaneity. *Journal of Marketing, 44,* 65–77.

Bagozzi, R. P., Fornell, C., & Larcker, D. F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research, 16,* 437–454.

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18–20.

Borgen, F. H., & Seling, M. J. (1978). Uses of discriminant analysis following MANOVA: Multivariate statistics for multivariate purposes. *Journal of Applied Psychology, 63,* 689–697.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practices, 20*(4), 6–18.

Campbell, D. T., & Erlebacher, A. (1975). How regression artifacts in quasiexperimental evaluations can mistakenly make compensatory education look harmful. In M. Guttentag & E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 1, pp. 597–617). Beverly Hills, CA: Sage.

Chambless, D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7–18.

Chandler, R. (1957). The statistical concepts of confidence and significance. *Psychological Bulletin, 54*, 429–430.

Cliff, N. (1987). *Analyzing multivariate data*. San Diego, CA: Harcourt Brace Jovanovich.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426–443.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement, 61*, 229–248.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532–575.

Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic Press.

Dunlap, W. P., & Landis, R. S. (1998). Interpretations of multiple regression borrowed from factor analysis and canonical correlation. *The Journal of General Psychology, 125*, 397–407.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An *EPM* guidelines editorial. *Educational and Psychological Measurement, 61*, 517–531.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8), 4–14.

Fidler, F. (2002). The fifth edition of the APA *Publication Manual*: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement, 62*, 749–770.

Fidler, F., Thomason, N., Cumming, G., Finch, S. & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but they can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*, 119–127.

Fish, L. J. (1988). Why multivariate methods are usually vital. *Measurement and Evaluation in Counseling and Development, 21*, 130–137.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: A reminder. *Structural Equation Modeling, 10*, 142–153.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hester, Y. C. (2000). An analysis of the use and misuse of ANOVA. (Doctoral dissertation, Texas A&M University, 2000). *Dissertation Abstracts International, 61*, 4332A. (UMI No. 9994257)

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley & Sons.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*, 227–240.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67*, 300–307.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183-220.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*, 239–251.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409–426.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43*, 443–477.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

Kendall, P. C. (1999). Clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 283–284.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart & Winston.

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research, 68*, 350–386.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69*, 280–309.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759.

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85*, 410–416.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in medical research. *The American Statistician, 52*, 127–132.

McLean, J. E., & Kaufman, A. S. (2000). Editorial: Statistical significance testing and other changes to *Research in the Schools, 7*(2), 1–2.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institution Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241–286.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart & Winston.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM* (2nd ed.). New York: Churchill Livingstone.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*, 115–129.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61*, 605–632.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 99–105). Greenwich, CT: JAI Press.

Snyder, P. (2000). Guidelines for reporting results of group quantitative investigations. *Journal of Early Intervention, 23*, 145–150.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*, 334–349.

Spearman, C. (1904). The proof and measurement of association between two things. *Journal of Psychology, 15*, 72–101.

Steiger, J. H., & Fouladi, R. T. (1992). R^2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers, 4*, 581–582.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation.* Newbury Park, CA: Sage.

Thompson, B. (1986). ANOVA versus regression analysis of ATI designs: An empirical investigation. *Educational and Psychological Measurement, 46*, 917–928.

Thompson, B. (1992). Misuse of ANCOVA and related "statistical control" procedures. *Reading Psychology, 13*, iii–xviii.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*, 525–534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.

Thompson, B. (1998a). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799–800.

Thompson, B. (1998b, July). *The ten commandments of good Structural Equation Modeling behavior: A user-friendly, introductory primer on SEM.* Paper presented at the annual meeting of the U.S. Department of Education, Office of Special Education Programs Project Directors' Conference, Washington, DC. (ERIC Document Reproduction Service No. ED 420 154)

Thompson, B. (1999a, April). *Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. (ERIC Document Reproduction Service No. ED 429 110)

Thompson, B. (1999b). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children, 65,* 329–337.

Thompson, B. (2000a). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 285–316). Washington, DC: American Psychological Association.

Thompson, B. (2000b). Ten commandments of structural equation modeling. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261–284). Washington, DC: American Psychological Association.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70,* 80–93.

Thompson, B. (2002a). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80,* 64–71.

Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24–31.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues.* Newbury Park, CA: Sage.

Thompson, B. (in press-a). Research synthesis: Effect sizes. In G. Camilli, P. B. Elmore, & J. Green (Eds.), *Complementary methods for research in education.* Washington, DC: American Educational Research Association.

Thompson, B. (in press-b). The "significance" crisis in psychology and education. *Journal of Socio-Economics.*

Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. *Educational and Psychological Measurement, 45,* 203–209.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6–20.

Vacha-Haase, T., Henson, R. K., & Caruso, J. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62,* 562–569.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509–522.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10,* 413–425.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58,* 21–37.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. [reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html]

Wright, S. (1921). Correlation and causality. *Journal of Agricultural Research, 20,* 557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5,* 161–215

ABOUT THE AUTHORS

BRUCE THOMPSON (CEC TX Federation), Texas A & M University and Baylor College of Medicine, College Station, Texas. **KAREN E. DIAMOND** (CEC IN Federation), Professor and Director, Child Development Laboratory School, Purdue University, W. Lafayette, Indiana. **ROBIN MCWILLIAM,** Director, Center for Child Development, Professor of Pediatrics and Special Education, Vanderbilt University Medical Center, Nashville, Tennessee. **PATRICIA SNYDER** (CEC #514), Associate Dean for Research and Graduate Studies, Louisiana State University Health Science Center, New Orleans. **SCOTT W. SNYDER** (CEC #144), School of Education Dean's Office, University of Alabama, Birmingham.

Correspondence concerning this article should be addressed to Bruce Thompson, TAMU Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225, or via e-mail using the Internet URL: http://www.coe.tamu.edu/-bthompson.

Manuscript received January 2004; accepted April 2004.