Current Issues in Education

Mary Lou Fulton College of Education
Arizona State University

Volume 6, Number 2 February, 2003 ISSN 1099-839X

Typology of Analytical and Interpretational Errors in Quantitative and Qualitative Educational Research

Anthony J. Onwuegbuzie
Howard University

Larry G. Daniel
University of North Florida

The purpose of this paper is to identify and to discuss major analytical and interpretational errors that occur regularly in quantitative and qualitative educational research. A comprehensive review of the literature discussing various problems was conducted. With respect to quantitative data analyses, common analytical and interpretational misconceptions are presented for data-analytic techniques representing each major member of the general linear model, including hierarchical linear modeling. Common errors associated with many of these approaches include (a) no evidence provided that statistical assumptions were checked; (b) no power/sample size considerations discussed; (c) inappropriate treatment of multivariate data; (d) use of stepwise procedures; (e) failure to report reliability indices for either previous or present samples; (f) no control for Type I error rate; and (g) failure to report effect sizes. With respect to qualitative research studies, the most common errors are failure to provide evidence for judging the dependability (i.e., reliability) and credibility (i.e., validity) of findings, generalizing findings beyond the sample, and failure to estimate and to interpret effect sizes.

Educational research worldwide has played a major role in influencing and informing educational practice. Indeed, the last decade has seen a proliferation in the number of articles published in educational research journals. Some of these published works have been the basis of educational reform in many settings. Moreover, most investigators utilize previous research in developing their conceptual and theoretical frameworks, as well as in providing qualitative (e.g., content analyses) and quantitative (e.g., meta-analyses) reviews of the literature in which the key findings are summarized. In conducting literature reviews, researchers often assume that the documented findings are trustworthy.

Unfortunately, some researchers have found that the majority of published studies and dissertations are seriously flawed, containing analytical and interpretational errors (Daniel, 1998c; Hall, Ward, & Comer, 1988; Keselman et al., 1998; Onwuegbuzie, 2002b; Thompson, 1998a; Vockell & Asher, 1974; Ward, Hall, & Schramm, 1975; Witta & Daniel, 1998). Some of these flaws have arisen from graduate-level instruction in which research methodology and statistics are taught as a series of routine steps, rather than as a holistic, reflective,

integrative process (Kerlinger, 1960; Newman & Benz, 1998); from graduate-level curricula that minimize students' exposure to quantitative and qualitative content (Aiken et al., 1990; Thompson, 1998a); from proliferations of various erroneous "mythologies" about the nature of research (Daniel, 1997; Kerlinger, 1960); from increasing numbers of research methodology instructors teaching out of their specialty areas; and from a failure, unwillingness, or even refusal to recognize that analytical and interpretational techniques that were popular in previous decades no longer reflect best practices and, moreover, may now be deemed inappropriate, invalid, or obsolete (Schmidt & Hunter, 1997).

Purpose

The purpose of the present paper is to identify and to discuss the major analytical and interpretational errors that occur in qualitative and quantitative educational research. Also contained in this essay are recommendations for good data analytic practices for each of these techniques that are based on the extant literature. Although a few methodologists have identified common errors made by researchers when analyzing various types of data, they have tended to focus their attention on a few data-analytic techniques at a time. For example, Thompson (1998a) reviewed the following five methodological errors that occur in educational research: (a) use of stepwise methods; (b) failure to consider result interpretation in the context specificity of analytical weights (e.g., regression beta weights, discriminant function coefficients); (c) failure to interpret both weights and structure coefficients in result interpretation; (d) failure to realize that reliability is a function of scores and not of instruments; and (e) incorrect interpretation of statistical significance and the associated failure to report and to interpret effect sizes present in all quantitative analyses. Moreover, no published paper was found in which errors made in both quantitative and qualitative research are discussed within the same article. As such, the present essay appears to represent the most wide-ranging discussion of analytical and interpretational errors in educational research to date.

There is little doubt that a myriad of analytical errors still prevails, despite the widespread availability of statistical software and many documented accounts calling for better research practice. As supported in the foregoing literature review, it is likely that many of these errors stem from a lack of a unified framework for analyzing and interpreting qualitative and quantitative data. In order to rectify these shortcomings, it is recommended that researchers be provided with a catalog of appropriate and inappropriate data analytic practices upon which the majority of researchers agree. In addition, editors of research journals should provide authors, as well as members of their editorial board and other reviewers of articles, with specific guidelines for analyzing and interpreting empirical data (Daniel, 1998a). Again, these guidelines should be based on agreed-upon practices. As such, the guidelines that are included in the ensuing literature review represent one small step in this direction..

Method for Arriving at Typology of Errors

Over the last six years, several methodologists have examined various erroneous analytical practices undertaken by applied quantitative researchers in various educational and psychological journals. The current authors attempted to obtain as many of these articles arising from these examinations as possible by conducting a search of literature databases such as ERIC (i.e., Educational Resource Information Center) and PsycINFO, as well as by attending methodological paper presentations at state (e.g., Georgia Educational Research Association), regional (e.g., Mid-South Educational Research Association, Southwest Educational Research Association, Eastern Educational Research Association, Midwestern Educational Research Association), national (e.g., American Educational Research Association, Association for the Advancement of Educational Research), and international (e.g., European Educational Research Association) conferences over this time period. The articles extracted from these sources included the following: Elmore and Woehlke (1988, 1998); Kirk (1996); Keselman et al. (1998); Onwuegbuzie (2002b), Vacha-Haase, Ness, Nilsson, and Reetz (1999); Vacha-Haase (1998); Simmelink and Vacha-Haase (1999); Henson, Capraro, and Capraro (2001); Henson and Roberts (in press); Lance and Vacha-Haase (1998); Vacha-Haase, Nilsson, Reetz, Lance, and Thompson (2000); Snyder and Thompson (1998); Thompson (1999b); Thompson and Snyder (1997, 1998), Vacha-Haase and Ness (1999); Vacha-Haase and Nilsson (1998); and McMillan, Lawson, Lewis, and Snyder (2002). All of these articles represent paper presentations at professional meetings and/or published works in reputable journals over the last six years.

In addition to searching the literature database and collecting methodological articles from professional meetings, the authors used the "snowballing" approach to obtaining manuscripts. Specifically, the reference list of every methodological paper extracted was examined to determine if it contained relevant articles that we had overlooked. This technique led to the identification of several additional articles. This method also helped

us to validate our choice of articles. For example, Vacha-Haase et al. (2000) reviewed 10 of the articles cited above, whereas McMillan et al. (2002) reviewed these same 10 articles, as well as Vacha-Haase et al.'s (2000) article. These three techniques for extracting methodological papers (i.e., database searching, attending conferences, snowballing) led to the identification of a large proportion of empirical studies examining the erroneous practices undertaken by applied researchers with respect to statistical analyses over the last six years.

The same three techniques outlined above were used to obtain articles that formed the basis of our recommendations for analyzing and interpreting quantitative and qualitative data. Extracting literature for this purpose was much more challenging than obtaining articles that examined the incidence of errors of omission and commission made by applied researchers, as described above. Whereas the latter type contained less than two dozen articles, papers presenting methodological recommendations for qualitative and statistical analyses numbered in the hundreds, just in the last decade alone! In extracting articles from the literature database, professional conferences, and via snowballing, the present writers tended to include articles that were authored or co-authored by quantitative and qualitative researchers and methodologists with national/international reputations. Many of these methodologists are not only researchers and writers, but also are journal editors and reviewers. Thus, they are widely read in the fields of educational and psychological research, as reflected by the reference lists contained in their articles. Snowballing techniques on these lead methodologists' articles yielded many more useful sources. In compiling our list of recommendations, we also reviewed many of the most popular textbooks in the area of qualitative research methods, quantitative research methods, statistics, measurement, and evaluation.

A series of content analyses was undertaken on the collected articles. Specifically, a content analysis was undertaken with respect to each of the analytical techniques discussed below. In using this procedure, our goal was to summarize the collective thought in the field. It could be argued that the fact that we did not summarize a random (i.e., scientific) sample of methodological papers provides a limitation to our paper. To the extent that our sample of articles was not representative of the recommendations posited by the majority of the leading methodologists, this criticism is valid. However, it should be noted that the aim of this essay was not to provide a survey of different analytical techniques found in the literature because this would have led to the "best" and the "worst" recommendations being given equal weight; rather, our goal was to attempt to determine the best practices as advanced by the community of research scholars as a whole. In any case, a perusal of other articles similar to our own (e.g., Thompson, 1994a, 1998a, 1999) indicates no more, and often even less structure in the technique used to select articles than described above. At the very least, as noted earlier, our paper appears to cite more literature in general and more current articles in particular than any other paper of its type.

Review of the Literature

Errors Common to Both Qualitative and Quantitative Research

At the highest level, analytical and interpretational errors in educational research include creating a false dichotomy between quantitative and qualitative research methodologies; that is, failing to treat quantitative and qualitative research strategies as lying on an interactive continuum, with theory as the driving force. This practice tends to prevent researchers from taking a holistic and comprehensive approach to research (Newman & Benz, 1998).

Throughout the 20th century, an uncompromising rift has prevailed between quantitative and qualitative researchers. Quantitative purists express assumptions about the world in general and research in particular that are consistent with a positivist or empiricist philosophy, whereas qualitative purists (e.g., post-positivists, post-structuralists, and post-modernists) reject positivism (Onwuegbuzie, 2002a). Moreover, the major differences that prevail between the two sets of purists are at the level of logic of justification (Smith & Heshusius, 1986). Positivists believe that behavior can be measured empirically. On the other hand, non-positivism is rooted in the constructivist, hermeneutic paradigm (i.e., *Verstehen*) in which multiple realities are socially constructed through individual and collaborative definitions of the situation, that values are an essential component of the research process, and that facts are indistinguishable from values (Onwuegbuzie, 2000a, 2002a). As such, qualitative purists are skeptical about the utility of providing evidence of representation and legitimation (Onwuegbuzie, 2000b). Unfortunately, much of the quantitative-qualitative debate has been counterproductive, entailing a continual contest of polemics, which has tended to confuse rather than to illuminate, and to segregate rather than to unify educational researchers (Onwuegbuzie, 2002a). Indeed, this trend prompted Miles and Huberman (1984, p. 21) to declare, "epistemological purity doesn't get research done."

However, more and more researchers are realizing that no one paradigm is a hegemony in

educational research. Indeed, as concluded by Hammersley (1992), the primary dilemma facing both sets of purists is that their assumptions are self-refuting. With respect to positivists, their assertion of the verifiability principle is self-refuting because it is neither logical nor empirical, and thus lacks meaning. To be congruous with their epistemological underpinnings, extreme relativists (i.e., constructivists) must concede that their assertion that all truth is relative is itself only true in the relative sense; thus, in terms of other philosophical perspectives their claims may be false. Accordingly, relativism is both true and false (Hammersley, 1992). Moreover, to be consistent with their tenets, realists must treat the quantitative paradigm not only as being true by its own standards, but also as a reality that is as legitimate as is any other reality--in particular, the qualitative paradigm (Onwuegbuzie, 2002a). As such, a false dichotomy exists between the quantitative and qualitative research paradigms (Newman & Benz, 1998).

As asserted by Onwuegbuzie (2002a), recognizing these flaws in the logic of justification allows one to re-frame how research paradigms should be considered. As surmised by Newman and Benz (1998), instead of representing a dichotomy, positivist and non-positivist ideologies lie on an epistemological continuum. In fact, the myriad of dichotomies that are used to differentiate qualitative and quantitative research paradigms can be re-framed as lying on continua. These include realism versus idealism, foundational versus antifoundational, objective versus subjective, impersonal versus personal, and deductive reasoning versus inductive reasoning (Onwuegbuzie, 2002a). Such a re-conceptualization permits quantitative and qualitative researchers alike to focus more on research methodologies rather than on paradigmatic considerations (Onwuegbuzie, 2002a). Indeed, as contended by Smith and Heshusius (1986), there is no one-to-one correspondence between research paradigm and research methodology.

Paradigm-Specific Errors in Educational Research

The remainder of this paper provides a critical synthesis and review of the educational research literature, examining both the extant qualitative and quantitative body of literature as described above. The first component involves an identification and discussion of the most prevalent analytical and interpretational errors made in qualitative educational research. This component is organized into sections that discuss general analytical and interpretational errors made in qualitative research, regardless of which methodologies are being referenced. On the other hand, the second component is divided into two major parts. The first part, mirroring the qualitative component, discusses general analytical and interpretational errors made in quantitative research, irrespective of the underlying technique(s). The second part provides common analytical and interpretational misconceptions for each of the major data-analytic techniques, including: bivariate correlational analyses, multiple regression, analysis of variance, analysis of covariance, multivariate analysis of variance, multiple analysis of covariance, discriminant analysis, exploratory factor analysis, confirmatory factor analysis, and structural equation modeling, as well as hierarchical linear modeling. Because of the inclusion of this second section, the quantitative component is much longer than is the qualitative component. However, the discrepancy in length should not be interpreted to mean that one paradigm is more important than is the other or that one paradigm leads to better quality research than does the other.

Errors in Qualitative Research

Many data analytic and interpretational errors permeate qualitative research. Therefore, the section below highlights the most common and pervasive errors encountered in the literature. In addition to discussing general errors made in qualitative research, the authors would have liked to have outlined the most prevalent analytical and interpretational errors that have been found to occur for each of the major data-analytic techniques-similar to that undertaken for quantitative research. However, this goal was beyond the scope of the present article for the following reasons. First and foremost, there is not the same level of agreement among qualitative researchers concerning available data-analytic approaches as there is among quantitative researchers. Indeed, whereby specific terms have been given to very specific quantitative (i.e., statistical) analyses (e.g., t-test, analysis of variance, multiple regression), the interactive nature of qualitative data analysis renders it much more difficult to provide labels for each type of analysis. For example, what one qualitative researcher might refer to as the method of constant comparison (e.g., Lincoln & Guba, 1985), another researcher might call a thematic analysis (e.g., Boyatzis, 1998). Second, whereas quantitative data analysis typically represents a distinct stage in the research process (Onwuegbuzie, in press-a), qualitative data-analysis tends to be much more interactive, recursive, and iterative. More specifically, in qualitative research, the research design/data collection, data analysis, and data interpretation stages are often non-linear in nature, and it is not unusual for these three stages to be inseparable. As a result, only general analytical

and interpretational errors that have been found to occur in qualitative research are presented.

General errors in qualitative research. ***Failure to legitimize research findings.*** With respect to qualitative research methodologies, analytical errors include a failure, often for philosophical reasons, to legitimize research findings and interpretations through documentation of validity (e.g., credibility, relativism, external criticism) and reliability (e.g., inter-rater reliability, internal consistency). With respect to the former, although the importance of validity has long been recognized by quantitative researchers, this issue has been the subject of disagreement among qualitative researchers. At the one end of the qualitative continuum are those (e.g., Goetz & LeCompte, 1984; Miles & Huberman, 1984) who contend that validity for qualitative research should be interpreted in the same manner as for quantitative research. At the other end of the continuum, some post-modernists (e.g., Wolcott, 1990) question the appropriateness of validity in qualitative research, asserting that the goal of providing evidences of validity is utopian. Disturbingly, a common definition of validity among relativists is that it represents whatever the community agrees it should represent. Unfortunately, such a definition is ambiguous, and, consequently, does not help beginning qualitative researchers to design their studies and to assess the legitimacy and trustworthiness of their findings.

It appears that a reason for the rejection of validity by some qualitative researchers stems from their perceptions that the positivist definition and interpretation of validity serves as the yardstick against which all other standards are evaluated. Thus, these extremists believe that in order to reject positivism, they must reject validity (Onwuegbuzie, 2002a). However, this is tantamount to throwing out the baby with the bath water.

Unfortunately, many qualitative researchers adopt an "anything goes" relativist attitude (Onwuegbuzie, 2002a), culminating in a failure to assess the credibility of their data interpretations. Yet, as contended by Onwuegbuzie (2002a), in order to be taken seriously, qualitative researchers must be accountable fully at all phases of their research study, including the data collection, analysis, and interpretation stages. Such accountability can only come to the fore by providing evidence of representation and legitimation.

Thus, rigor in research is needed, regardless of whether quantitative or qualitative research techniques are utilized. With respect to the latter, it is important that qualitative researchers assess the legitimacy of their interpretations. This can be undertaken by re-defining the concept of validity in qualitative research, for example, by deeming validity as representing an examination of rival interpretations and hypotheses (Polkinghorne, 1983), or by re-conceptualizing validity as being multi-dimensional (e.g., credibility, transferability, dependability, confirmability; Lincoln & Guba, 1985). In fact, Onwuegbuzie (2000b) identified 24 techniques for assessing the legitimacy of qualitative findings and interpretations.

A myriad of methods for assessing the truth value of findings and interpretations in qualitative research have emerged in the literature. These include triangulation, prolonged engagement, persistent observation, leaving an audit trail, member checking, weighting the evidence, checking for representativeness of sources of data, checking for researcher effects, making contrasts/comparisons, checking the meaning of outliers, using extreme cases, ruling out spurious relations, replicating a finding, assessing rival explanations, looking for negative evidence, obtaining feedback from informants, peer debriefing, clarifying researcher bias, and thick description (Creswell, 1998; Maxwell, 1996; Miles & Huberman, 1984, 1994). Utilizing and documenting such techniques should help to reduce methodological errors in qualitative research.

Another analytical error in qualitative research is the failure to assess the reliability of observed findings. Although reliability is generally conceived of as an important concept in analyzing quantitative data, it is important to note that reliability is actually pertinent with regard to qualitative data (Madill, Jordan, & Shirley, 2000). In qualitative research, information gleaned from observations, interviews, and the like must be "trustworthy" (Eisenhart & Howe, 1992; Lincoln & Guba, 1985); otherwise any themes that emerge from these data will not be credible. An important component of trustworthiness is "dependability" (Lincoln & Guba, 1985). Interestingly, dependability is analogous to reliability (Eisenhart & Howe, 1992). Techniques for evaluating this dimension of trustworthiness include triangulation, which involves the use of multiple and different methods, investigators, sources, and theories to obtain corroborating evidence (Ely, Anzul, Friedman, Garner, & Steinmetz, 1991; Glesne & Peshkin, 1992; Lincoln & Guba, 1985; Merriam, 1988; Miles & Huberman, 1984, 1994; Patton, 1990). Triangulation reduces the possibility of chance associations, as well as of systematic biases prevailing due to only qualitative (or quantitative) methods being utilized, thereby allowing greater confidence in any interpretations made (Fielding & Fielding, 1986; Maxwell, 1992).

According to Denzin (1978), three outcomes arise from triangulation: convergence, inconsistency, and contradiction. Each of these outcomes clearly represents issues pertaining to reliability. Nevertheless, many interpretivists refrain from using the term "reliability" when pertaining to qualitative data, probably because of an attempt to distance qualitative analytical techniques from statistical method (Madill et al., 2000). However, this line of thinking is counterproductive. Indeed, as noted by Constas (1992, p. 255), unless methods for examining rival hypotheses in qualitative research are developed, "the research community will be entitled to question the analytical rigor of qualitative research"--where rigor is defined as the attempt to make data and categorical schemes as public and as replicable as possible (Denzin, 1978). Fortunately, reliability as a concept in qualitative data analysis is increasingly gaining acceptance. In particular, it is no longer unusual for qualitative researchers to report either intrarater (e.g., consistency of a given rater's scores or observations) or interrater (e.g., consistency of two or more independent raters' scores or observations) reliability estimates (Gay & Airasian, 2000; Worthen, Borg, & White, 1993). Evidence of this can be gleaned from the fact that a leading theory-building qualitative software program called NUD.IST (non-numerical unstructured data indexing searching & theorizing) allows data analysts to determine inter-coder reliability (QSR International Pty Ltd., 2002).

Generalizing findings beyond sample. Interpretative errors in qualitative research include the tendency to generalize findings rather than to use qualitative techniques to obtain insights into particular educational, social, and familial processes and practices that existed within a specific location (Connolly, 1998). Only when relatively large representative samples are utilized is it fully justified for researchers to generalize findings from the sample to the population. While obtaining large, representative samples typically is the goal in quantitative research, this is not the case in qualitative research, where purposive sampling of relatively few cases is more the norm. Yet, some qualitative researchers find it difficult to resist the temptation to generalize their results (e.g., thematic representations) to the underlying population.

Failure to estimate and interpret effect sizes. Recently, the American Psychological Association (APA) Task Force advocated strongly that researchers should "always present effect sizes for primary outcomes...[and]...reporting and interpreting effect sizes...is essential to good research" (Wilkinson & the Task Force on Statistical Inference, 1999, pp. 10-11). However, the title of their report (i.e., "Statistical Methods in Psychology Journals: Guidelines and Explanations"), suggests that these stipulations are pertinent only to *quantitative* data. Moreover, the APA Task force did not provide any recommendations that effect sizes be reported and interpreted when analyzing qualitative data. Yet, as advanced by Onwuegbuzie (in press-b), there are many situations in which effect sizes would provide a richer, thicker description of underlying *qualitative* data. Indeed, it appears that failure to utilize effect sizes by qualitative researchers stems, at least in part, from educational researchers associating effect sizes with the quantitative paradigm. Yet, ironically, the use of effect sizes actually results in empirical data being *qualitized* (Tashakkori & Teddlie, 1998), which, in turn, facilitates the assessment of whether an observed effect is small, medium, large, or the like (Cohen, 1988). Simply put, these effect size interpretations represent *qualitative* categorizations (Onwuegbuzie & Teddlie, 2002).

At its most basic form, providing an effect size in qualitative research involves obtaining counts of the frequency of an observed phenomenon. Interestingly, as noted by Sechrest and Sidani (1995, p. 79), "qualitative researchers regularly use terms like 'many,' 'most,' 'frequently,' 'several,' 'never,' and so on. These terms are fundamentally quantitative." Moreover, it could be argued that terms such as "many" are "frequently" are relative; that is, they depend on the context from which the data were obtained. Using such phrases without supplementing them with the counts, forces the reader to accept the writer's interpretation. Conversely, by providing counts, readers can make up their own mind as to what adjective best depicts the underlying phenomenon. As a result, qualitative researchers often can extract more meaning by obtaining counts of observations in addition to their narrative descriptions (Sandelowski, 2001).

For example, Witcher, Onwuegbuzie, and Minor (2001) conducted a qualitative study to ascertain preservice teachers' perceptions of characteristics of effective teachers. A phenomenological analysis resulted in the emergence of six characteristics of effective teaching (as perceived by the preservice teachers). By counting the frequency of the emergent themes, these researchers found that of the six identified characteristics of effective teachers, student-centeredness was the most commonly-cited trait (cited by 80% of the preservice sample). This was followed by enthusiasm for teaching (40%), ethicalness (39%), classroom and behavior management (33%), teaching methodology (32%), and knowledge of subject (32%). This example provides support for Dey's (1993) contention that

meaning and number can be inextricably intertwined. Obtaining counts of the themes prevented the researchers from over-weighting or under-weighting the emergent themes (Sandelowski, 2001).

The development of themes, categories, typologies, and the like is commonplace in qualitative data analysis (Boyatzis, 1998; Constas, 1992). Such development is based on the frequency with which a facet occurs (Miles & Huberman, 1994). More specifically, every time a qualitative researcher reduces data to categories or themes, he/she is utilizing the "numbered nature of phenomena for their analysis" (Sandelowski, 2001, p. 231). In fact, at least three rationales prevail for counting themes: (a) to identify patterns more easily, (b) to maintain analytic integrity, and (c) to verify a hypothesis (Miles & Huberman, 1994). Further, by adding numerical accuracy to their descriptive accounts, Witcher et al. (2001) were able to leave an audit trail, which involved a more extensive documentation of the observed data. Interestingly, audit trails are advocated by qualitative researchers as a means of evaluating legitimation or increasing legitimation, or both (Halpern, 1983; Lincoln & Guba, 1985).

Counting themes is a manifestation of what Tashakkori and Teddlie (1998, p. 126) referred to as "quantitizing" data, in which qualitative data are transformed into numerical codes that can be represented statistically. As stated by Sandelowski (2001), in quantitizing, "qualitative 'themes' are numerically represented, in scores, scales, or clusters, in order more fully to describe and/or interpret a target phenomenon" (p. 231). Also, Boyatzis (1998, p. 129) referred to the counting of themes as "quantitative translation."

Onwuegbuzie (in press-b) presented a typology of effect sizes in qualitative research. This typology was divided into what he termed *manifest effect sizes* (i.e., effect sizes pertaining to observable content) and *latent effect sizes* (i.e., effect sizes pertaining to non-observable, underlying aspects of the phenomenon under observation). For example, when conducting thematic analyses, qualitative analysts usually only classify and describe emergent themes. However, much more information can be ascertained about these themes. In particular, these themes can be quantitized (i.e., quantified) by determining the frequency of occurrence (e.g., least/most prevalent theme) and intensity of each identified theme (Onwuegbuzie, in press-b). Moreover, by unitizing the themes and utilizing what he termed as *intra-respondent matrices* (i.e., unit x theme matrices) and *inter-respondent matrices (e.g., subject x theme matrices and subject x unit matrices)*, Onwuegbuzie demonstrated how exploratory factor analyses and cluster analyses can be undertaken on these matrices such that the hierarchical structure of the themes (i.e., *meta-themes*) and their inter-relationships can be identified. Onwuegbuzie also illustrated how effect sizes (e.g., eigenvalues and proportion of variance explained by each theme) pertaining to the thematic structure and relationships among themes and meta-themes can be estimated.

Onwuegbuzie (in press-b) introduced the concept of *adjusted effect sizes* in qualitative research, in which the frequency and intensity of themes are adjusted for the time occurrence and length of the unit of analysis (e.g., observation, interview, text). For instance, with regard to the length of unit analysis, the frequency of the emergent theme could be divided by the number of words, sentences, paragraphs, and/or pages analyzed. Such adjusted effect sizes help to minimize bias that is inherent in the data (Onwuegbuzie, in press-b).

Consistent with Onwuegbuzie's conceptualization of effect sizes in qualitative research, nearly one-half a century ago, Barton and Lazarsfeld (1955) advocated the use of what they coined as "quasi-statistics" in qualitative research. According to these authors, quasi-statistics refer to the use of descriptive statistics that can be extracted from qualitative data. Interestingly, Becker (1970, pp. 81-82) contended that "one of the greatest faults in most observational case studies has been their failure to make explicit the quasi-statistical basis of their conclusions." As noted by Maxwell (1996):

> Quasi-statistics not only allow you to test and support claims that are inherently quantitative, but also enable you to assess the *amount* of evidence in your data that bears on a particular conclusion or threat, such as how many discrepant instances exist and from how many different sources they were obtained. (p. 95) [emphasis in original]

Indeed, Becker, Geer, Hughes, and Strauss (1961/1977) provided more than 50 tables and graphs in their qualitative work. These tables and graphs facilitate effect size interpretations of their qualitative data.

Errors in Quantitative Research

As is the case for qualitative research, there are many data analytic and interpretational errors that prevail in existing research that uses quantitative data regardless of the statistical analysis used. Thus, the first part of this section provides a summary of the major errors that are not dependent on the statistical technique used. In the second part of this section, we outline the errors in quantitative research that are dependent, at least for the most part, on the method used. Specifically, we present the major analytical and interpretational errors that have been found to

prevail for each of the major data-analytic techniques, namely, bivariate correlational analyses, reliability analyses, analysis of variance, analysis of covariance, multiple regression, multivariate analysis of variance, multiple analysis of covariance, discriminant analysis, canonical correlation analysis, principal component and factor analysis, confirmatory factor analysis, path analysis, structural equation modeling, and hierarchical linear modeling.

General errors in quantitative research. As noted by Onwuegbuzie (in press-a), threats to internal validity in quantitative research occur at both the data analysis (i.e., analytical errors) and data interpretation (i.e., interpretational errors) stages. Indeed, Onwuegbuzie (in press-a) described several types of errors that occur at both these stages. According to this researcher, data analytical errors can stem from several sources, including the following: mortality, non-interaction seeking bias, researcher bias, treatment replication error, violated assumptions, multicollinearity, and mis-specification error. Each of these sources of error is summarized briefly below. (For a more detailed discussion of these sources of error, see Onwuegbuzie, in press-a.)

Mortality. It is not uncommon for researchers to delete some cases from their final data sets. There are many reasons why such a practice occurs. Specifically, cases may be deleted if they appear to represent outlying observations. Alternatively, the size of a data set may be reduced in an attempt to analyze groups with equal or approximately equal sample sizes (i.e., to conduct a "balanced" analysis). Such removal of cases can lead to analytical errors in the first situation if one or more participants who are deleted represent valid cases, and in the second situation if one or more participants who are removed from the data set are different than those who remain. In either event, the reduction of the data set introduces or adds bias into the analysis, thereby influencing the effect size in an unknown manner (Onwuegbuzie, in press-a). In the same way, using casewise deletion and listwise deletion strategies in the presence of missing data, a very common practice among researchers, also can lead to analytical errors.

Non-interaction seeking bias. When testing hypotheses and theory, some researchers do not examine the presence of interactions. This likely is more often to occur for correlational-based analyses (e.g., correlations, regression, canonical correlation, path analysis, structural equation modeling) than for OVA-type methods (e.g., factorial analysis of variance, multivariate analysis of variance). Non-interaction seeking bias can not only lead to errors at the data analysis stage, but it can also induce interpretational errors. Moreover, by not formally testing for the presence of interactions, researchers may end up interpreting a model that does not accurately or validly represent the underlying nature of reality (Onwuegbuzie, in press-a).

Researcher bias. The form of researcher bias that is more prevalent at the data analysis stage is the halo effect. The halo effect occurs when a researcher is scoring open-ended responses, or the like, and allows her or his prior knowledge of or experience with the participants to influence the scores given. This biases the data, leading to analytical errors.

Treatment replication error. As noted by McMillan (1999), a common mistake made by analysts involves the use of an inappropriate unit of analysis. For example, a researcher might use individuals as the unit of analysis to compare groups when analyzing available group scores would have been more appropriate. In particular, analyzing individual data when groups received the intervention violates the independence assumption, thereby inducing an analytical error through the inflation of both the Type I error rate and effect size estimates.

Violated assumptions. Several authors (e.g., Keselman et al., 1998; Onwuegbuzie, 2002b) have noted that the majority of researchers do not adequately check the underlying assumptions associated with a particular statistical test. Regardless of the inferential statistical technique used, unless assumptions are checked, the extent to which an analytical error prevails is unknown. With knowledge of the extent to which assumptions are violated, researchers are in a position to interpret findings within an appropriate context. However, when it is unknown whether assumptions have been met, data interpretation can be extremely misleading and invalid.

Multicollinearity. Multicollinearity occurs when two or more independent variables are highly related. When one independent variable is perfectly correlated with other independent variables, the parameter estimates are not uniquely determined. A strong, but less-than-perfect, linear relationship among independent variables, as is more often the case, results in unstable (least-squares) coefficients with large standard errors and wide confidence intervals (Fox, 1997). Multicollinearity often is associated with multiple regression; however, multicollinearity is an issue for other members of the general linear model, including OVA methods. Thus, multicollinearity should not only be assessed when multiple regression is involved but for all analysis involving two or more independent variables.

Mis-specification error. Mis-specification error involves omitting one or more important variables from the final model. This is an error that

can be committed with any inferential analysis. As noted by Onwuegbuzie (in press-a), mis-specification error often arises from a weak or non-existent theoretical framework for building a statistical model. This inattention to a theoretical framework leads many researchers to (a) undertake univariate analyses when the phenomenon is multivariate, (b) utilize data-driven techniques such as stepwise multiple regression procedures, and (c) omit the assessment of interactions. All of these approaches lead to mis-specification error. Unfortunately, mis-specification error, although likely common, is extremely difficult to detect, especially if the selected *non-optimal model*, which does not include any interaction terms, appears to fit the data adequately.

As noted by Onwuegbuzie (in press-a), interpretational errors, can arise from the following: effect size, confirmation bias, distorted graphics, illusory correlation, crud factor, positive manifold, and causal error. Each of these sources of error is summarized briefly below. (For a more detailed discussion of these sources of error, see Onwuegbuzie, in press-a.)

Effect size. The non-reporting of effect sizes likely represents the most common interpretational error in quantitative research. Failure to report effect sizes often culminates in misinterpretation of p-values. In particular, a p-value tends to be under-interpreted when the sample size is small and the corresponding non-reported effect size is large. On the other hand, a p-value tends to be over-interpreted when the sample size is large and the non-reported effect size is small (e.g., Daniel, 1998a). The lack of reporting of effect sizes led the APA Task Force to recommend strongly that researchers "always present effect sizes for primary outcomes...[and]...reporting and interpreting effect sizes...is essential to good research" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599). More recently, the latest version of the American Psychological Association (APA), version 5 (2001), contained the following statement:

> When reporting inferential statistics (e.g., *t* tests, *F* tests, and chi-square), include information about the obtained magnitude or value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme as or more extreme than the one obtained, and the direction of the effect. Be sure to include sufficient descriptive statistics (e.g., per-cell sample size, means, correlations, standard deviations) so that the nature of the effect being reported can be understood by the reader and for future meta-analyses. This information is important, even if no significant effect is being reported. (p. 22)

A few pages later, APA (2001) states

> Neither of the two types of probability value directly reflects the magnitude of an effect or the strength of a relationship. For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section. (p. 25)

On the next page, APA states that

> The general principle to be followed, however, 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. (p. 26)

Confirmation bias. Confirmation bias is the tendency for interpretations and conclusions based on the current data to be overly consistent with *a priori* hypotheses (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). Unfortunately, confirmation bias is a common source of error at the data interpretation stage. Confirmation bias tends to prevail when the researcher attempts to test theory, because testing a theory can "dominate research in a way that blinds the researcher to potentially informative observation" (Greenwald et al., 1986, p. 217). When hypotheses are not supported, some researchers interpret their data as if the theory underlying the hypotheses is still likely to be correct. In so doing, many researchers are not aware that the purpose of their research no longer can be described as theory testing but theory confirming.

However, confirmation bias, per se, does not necessarily lead to interpretational errors. Such errors occur only when one or more plausible rival explanations to underlying findings exist that could have been demonstrated as being superior if given the opportunity (Greenwald et al., 1986). However, because the vast majority of findings generate rival explanations, researchers should always assess their interpretations for the possibility of confirmation bias.

Distorted graphics. The interpretation of graphs can be a source of error. For example, it is not unusual for histograms to suggest normality, when numerical data (e.g., skewness and kurtosis) indicate non-normality. Distorted graphs could be the result of an inappropriate scale. Alternatively, interpretation errors can ensue even if the graph is not distorted, especially when the researcher has a confirmation bias (e.g., desperately wants to demonstrate that the normality assumption holds).

Illusory correlation. The illusory correlation is a propensity to overestimate relationships among variables that are either not related or only slightly related. The illusory correlation often arises from a

confirmation bias. The illusory correlation also may arise from a false consensus bias, in which researchers falsely believe that most other persons share their interpretations of a relationship (Johnson & Johnson, 2000).

Crud factor. Meehl (1990) observed that given a large enough sample size, many trivial relationships can emerge as being statistically significant because to some degree, "everything correlates to some extent with everything else" (p. 204). Meehl referred to this tendency to reject null hypotheses when the true relationships are trivial as the *crud* factor. This crud factor leads some researchers to interpret trivial relationships, leading to interpretational errors.

Positive manifold. Positive manifold can occur when individuals who perform well on one ability or attitudinal measure tend to perform well on other measures in the same domain (Neisser, 1998). As such, positive manifold can lead to an over-interpretation of relationships. Thus, analysts should be careful when interpreting relationships found among two or more sets of cognitive test scores or affective measures.

Causal error. Some researchers cannot resist interpreting large relationships as suggesting causality. However, causality is a function of the research design and not the analytic technique used. Thus, regardless of the complexity of the analysis, cause-and-effect relationships should only be inferred confidently from experimental studies. In the absence of an experimental design, any causal statements made likely will represent interpretational errors.

Additional general analytical and interpretational errors. Daniel and Onwuegbuzie (2000) have identified 10 analytical and interpretational errors associated with statistical significance testing. They labeled these errors as Type I to Type X. The first four errors are known to all statisticians as Type I (falsely rejecting the null hypothesis), Type II (incorrectly failing to reject the null hypothesis), Type III (incorrect inferences about result directionality), and Type IV (incorrectly following-up an interaction effect with a simple effects analysis). Daniel and Onwuegbuzie (2000) identified and described the following six additional types of error: (a) Type V error--internal replication error-measured via incidence of Type I or Type II errors detected during internal replication cycles when using methodologies such as the jackknife procedure; (b) Type VI error--reliability generalization error--measured via linkages of statistical results to characteristics of scores on the measures used to generate results (a particularly problematic type of error when researchers fail to consider differential reliability estimates for subsamples within a data set); (c) Type VII error--heterogeneity of variance/regression--measured via the extent to which data examined via analysis of variance/covariance are not appropriately screened to determine whether they meet homogeneity assumptions prior to analysis of group comparison statistics; (d) Type VIII error--test statistic distribution error--measured as the extent to which researchers express alternative hypotheses as directional yet evaluate results with two-tailed tests; (e) Type IX error--sampling bias error--measured via disparities in results generated from numerous convenience samples across a multiplicity of similar studies; and (f) Type X error--degrees of freedom error--measured as the tendency of researchers using certain statistical procedures (mainly stepwise procedures) erroneously to compute the degrees of freedom used in these methods.

Method-dependent errors in quantitative research. ***Correlation coefficients.*** With respect to quantitative research methodologies, perhaps the most common analytical/interpretational error stems from a failure to realize that all parametric analyses (i.e., univariate and multivariate techniques), with the exception of predictive discriminant analyses, are subsumed by a general linear model (GLM), and that, consequently, all analyses are correlational (Cohen, 1968; Henson, 2000; Knapp, 1978; Roberts & Henson, 2002; Thompson, 1998a). In particular, many researchers are unaware that even correlation coefficients are specific cases of the GLM, and are therefore bounded by its assumptions (Onwuegbuzie & Daniel, 2002a). Moreover, Onwuegbuzie and Daniel (2002a) identified several inappropriate practices undertaken by researchers while utilizing correlational coefficients for inferential purposes, including failure to consider the statistical assumptions underlying correlation coefficients, failure to interpret confidence intervals and effect sizes of correlation coefficients, failure to interpret p-calculated values in light of familywise Type I error, failure to consider the power of tests of hypotheses, failure to consider whether outliers are inherent in the data set, failure to recognize how measurement error can affect correlation coefficients, and failure to evaluate empirically the replicability of correlation coefficients (i.e., internal replication).

Based on these observations, Onwuegbuzie and Daniel (2002a) made the following 10 recommendations for utilizing and interpreting correlation coefficients:

1. Always check statistical assumptions *prior* to using Pearson's *r* to conduct tests of statistical significance, as well as after the correlation has been computed. Use non-

parametric correlation (e.g., Spearman's rho) if the normality assumption is violated.
2. Always adjust for Type I error when conducting multiple NHSTs [null hypothesis statistical tests] of correlations.
3. Always be cognizant of the power of NHSTs of correlations, preferably before the data collection stage, and, at the very least, at the data analysis stage.
4. When making inferences about the Pearson r value, always interpret effect sizes.
5. Do not conduct "nil" null tests of statistical significance for reliability and validity coefficients (i.e., do not test whether reliability and validity coefficients are statistically significantly greater than zero).
6. Do not report disattenuated correlation coefficients without also presenting the raw coefficients.
7. Do not correlate variables without a theoretical framework.
8. Avoid inferring causation from a correlation coefficient, regardless of its magnitude.
9. Do not use Hotelling's *t*-test when comparing correlation coefficients arising from the same sample.
10. Conduct external replications when possible, and, in their absence, always undertake internal replications.

Regarding recommendation (10) above, Onwuegbuzie and Daniel (2002a) coined the term "Type V error" to describe internal replication error rates (as noted earlier), which provides information about how stable the computed *p*-value is across multiple re-samples of the same dataset. Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias, violated assumptions, and mis-specification error are analytical errors that are particularly pertinent for correlation coefficients, whereas effect size, confirmation bias, illusory correlation, crud factor, positive manifold, and causal error are pertinent interpretational errors.

Reliability of scores. Authors of statistics textbooks routinely report that statistical power is affected by at least three components: (a) sample size, (b) level of statistical significance, and (c) effect size. However, a fourth component should be added, namely, the reliability of scores. Reliability, which typically ranges from 0 (measurement is all error) to 1 (no error in measurement), is the proportion of variance in the observed scores which is free from error. (Reliability coefficients also can be negative.)

Unfortunately, relatively few researchers report reliability coefficients for data from their samples (Meier & Davis, 1990; Onwuegbuzie, 2002b; Onwuegbuzie & Daniel, 2002a, 2002b; Thompson & Snyder, 1998; Vacha-Haase et al., 1999; Willson, 1980). For example, Willson (1980) noted "That reliability ...is unreported in almost half of the published research...[and is] inexcusable at this late date" (pp. 8-9). Unfortunately, more than two decades later, Vacha-Haase et al. (1999), who reviewed practices regarding the reporting of reliability coefficients in three journals from 1990 to 1997, found that 64.4% of articles did not provide reliability coefficients for the data being analyzed. Similarly, Vacha-Haase (1998), who identified 628 articles in which the Bem Sex Role Inventory (Bem, 1981) was utilized, found that 86.9% of the articles did not present any score reliability information for the underlying data. Simmelink and Vacha-Haase (1999) reported that 75.9% fell into this category with respect to the use of the Rosenberg Self-Esteem Instrument (Rosenberg, 1965).

The trend of not reporting current-sample reliability coefficients stems, in part, from a failure to realize that reliability is a function of scores, not of instruments (Thompson & Vacha-Haase, 2000).The dearth in the reporting of reliability estimates led the APA Task Force on Statistical Inference recently to recommend that authors "provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 21). Further, the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA/APA/NCME] (1999) stated that good standards for reporting results necessitate researchers to provide reliability estimates and standard errors for "each total score, subscore, or combination of scores that is to be interpreted" (p. 31).

Without information about score reliability, it is impossible to assess accurately the extent to which statistical power is affected. Thus, reliability coefficients always should be reported for the underlying data. Moreover, the use of confidence intervals around reliability coefficients is advocated, considering that reliability coefficients represent only point estimates. In fact, confidence intervals around reliability coefficients can be compared to coefficients presented in test manuals to assess generalizability (Onwuegbuzie & Daniel, 2002b).

Additionally, as noted by Onwuegbuzie and Daniel (2002b), the recommendations of the APA Task Force and AERA/APA/NCME (1999) regarding the reporting of current-sample reliability coefficients do not go far enough. Indeed, it is argued that reliability coefficients should not only be reported for the full sample at hand, but also for sample subgroups. For example, in a two-sample case, it is

possible to obtain a moderate reliability estimate for the full sample, whereby the reliability coefficient of one group is relatively large but the coefficient for the other group is relatively small. It is likely that such a case would produce a different outcome in terms of statistical and practical significance than would a scenario in which the ratio of reliability coefficients is much smaller. Simply put, comparing subgroups with different reliability coefficients can affect Type I and Type II error rates, as well as effect size estimates. In such circumstances, Type VI error (reliability generalization error; Daniel & Onwuegbuzie, 2000) prevails. Thus, in summary, we recommend that subgroup reliability coefficients be reported whenever possible, alongside their confidence intervals.

When current-sample reliability coefficients are not available, researchers, at the very least, should compare the sample composition and variability of scores of the present sample with those of the inducted (i.e., norm) group (Vacha-Haase, Kogan, & Thompson, 2000). The results of these comparisons should be delineated. Specifically, as noted by Vacha-Haase et al. (2000), assuming that previously-reported reliability coefficients generalize to the present sample is only marginally justified if the compositions and the score variabilities of the two samples are similar. Additionally, Magnusson's (1967) formula could be used to predict the reliability of the present sample, based on the reliability of the inducted sample and the standard deviations of the

$$R_c = 1 - \frac{\sigma_i^2(1 - R_i)}{\sigma_c^2}$$

inducted and current samples, as follows: where R_c = the predicted reliability of the current sample, R_i = the predicted reliability of the inducted sample, σ_c^2 is the variance of the current sample, and σ_i^2 is the variance of the inducted sample. However, it should be noted that the predicted reliabilities are purely theoretical. (For an example of the use of this formula see Diamond & Onwuegbuzie, 2001.)

Independent/Dependent Samples t-test. When researchers are interested in comparing two independent samples, assuming normality, they must choose between the pooled and non-pooled *t*-test. This selection depends on whether the variances are equal or unequal, respectively. When the variances are equal, the pooled *t*-test should be used. On the other hand, when the variances are unequal, or when there is doubt about their equality, the non-pooled *t*-test should be employed. That is, if the homogeneity of variance assumption does not hold, then the t-test formula for separate variances should be used (Maxwell & Delaney, 1990). Under the assumption of variance homogeneity, the pooled *t*-test is only slightly more powerful (i.e., smaller Type II error probability) than is the non-pooled *t*-test. At the same time, in the presence of variance heterogeneity, use of the pooled *t*-test can increase greatly the chances of an invalid conclusion, especially when the sample sizes also are unequal.

Some statisticians recommend that the analyst first test the equal variance hypothesis:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ vs. } H_a: \sigma_1^2 \neq \sigma_2^2$$

If H_0 is rejected, then they conclude that the population variances are not equal and proceed with the non-pooled t-test. If H_0 is not rejected, then they advise that the pooled *t*-test procedure be used. However, this is a misuse of the variance test, because it may result in *failing to reject* the null hypothesis that $\sigma_1^2 = \sigma_2^2$ without knowing the probability of a Type II error. It should be noted that when the researcher fails to reject a null hypothesis, it is inappropriate to conclude that the null is true, but only that there is not enough evidence to justify its rejection. In addition, the probability of a Type I error is changed by performing two tests. Thus, we recommend the use of the pooled t-test only when prior knowledge, experience, or theory suggests that the population variances are approximately equal. If there is any doubt about the equality of the variances, the non-pooled *t*-test should be used. Unfortunately, presently, it appears that many analysts utilize the pooled version of the *t*-test. Such a practice typically will result in Type VII error (heterogeneity of variance/regression; Daniel & Onwuegbuzie, 2000). Disturbingly, Keselman et al. (1998) found that variance homogeneity was evaluated in only 8.20% of articles involving between-subjects univariate designs (n = 61) published in 17 prominent educational and behavioral science research journals in the 1994 or 1995 issues that were selected for review. This trend must be reversed in order to strengthen the conclusion validity stemming from independent/dependent t-tests. Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias, violated assumptions, and mis-specification error are analytical errors that are particularly pertinent for independent/dependent samples *t*-tests, whereas effect size, confirmation bias, and causal error are common interpretational errors.

Analysis of variance tests. Research suggests that analysis of variance (ANOVA) is the most popular statistical procedure for conducting null hypothesis statistical significance tests among educational researchers (Elmore & Woehlke, 1998; Goodwin & Goodwin, 1985; Onwuegbuzie, 2002b). Unfortunately, the ANOVA test is often misused. Specifically, lacking the knowledge that nearly all

parametric analyses represent the general linear model, many analysts inappropriately categorize variables in non-experimental designs using ANOVA, in an attempt to justify making causal inferences, when all that occurred typically is a discarding of relevant variance (Cliff, 1987; Pedhazur, 1982; Prosser, 1990; Thompson, 1986, 1988a, 1992a). For example, Cohen (1983) calculated that the Pearson product-moment correlation between a variable and its dichotomized version (i.e., divided at the mean) was .798, which suggests that the cost of dichotomization is approximately a 20% reduction in correlation coefficient. In other words, an artificially dichotomized variable accounts for only 63.7% as much variance as does the original continuous variable. Interestingly, more recently, Peet (1999) found that for the one-way ANOVA framework, as the number of categorized groups decreases (minimum number = 2), less variance in the dependent variable is accounted for by the categorical variable, compared to the continuous variable. Specifically, Peet noted that with four groups, almost 90% of the variance accounted for by the continuous variable was explained by the categorical variable; however, with two groups, only approximately 50% of the original variance accounted for was explained by the categorical variable. It follows that with factorial ANOVAs, when artificial categorization occurs, even more power is sacrificed. For instance, in the 2 x 2 ANOVA framework, when the correlation is between .2 and .5, double dichotomization at the mean culminates in a discarding of 60% of the sample members at both the two-tailed 5% and 1% levels. Thus, as stated by Kerlinger (1986), we recommend that researchers avoid artificially categorizing continuous variables, unless compelled to do so as a result of the distribution of the data (e.g., bimodal). Indeed, rather than categorizing independent variables, in many cases, regression techniques should be used, because they have been shown consistently to be superior to OVA methods (Daniel, 1989a; Kerlinger & Pedhazur, 1973; Lopez, 1989; Nelson & Zaichkowsky, 1979; Thompson, 1986).

Disturbingly, it is clear that most users of ANOVA tests do not adequately check the underlying assumptions. This is evidenced by the paucity of researchers who provide information about the extent to which ANOVA assumptions are met. For example, Keselman et al. (1998) reported that less than one-fifth of articles (i.e., 19.7%) "indicated some concern for distributional assumption violations" (p. 356). Similarly, Onwuegbuzie (2002b) found that only 11.1% of researchers discussed the extent to which OVA assumptions were violated. The fact that, when conducting univariate comparisons, ANOVA tests are almost exclusively used is a serious cause for concern, bearing in mind that (a) ANOVA relies on the assumptions of normality and homogeneity of variance; and (b) only the minority of data utilized in the field of educational research tends to satisfy ANOVA assumptions (Micceri, 1989; Wilcox, Charlin, & Thompson, 1986). Unfortunately, non-normality and variance heterogeneity lead to a distortion of Type I and/or Type II error rates, particularly if the group sizes are very different (Keselman et al., 1998). Thus, researchers always should check the ANOVA assumptions. In particular, if the normality assumption is violated, analysts should consider using the non-parametric counterparts, for example, the Mann-Whitney U test (for the two-group case) or the Kruskal-Wallis test (when three or more groups are being compared). When the homogeneity of variance assumption is violated (Type VII error; Daniel & Onwuegbuzie, 2000), techniques such as Welch, James, and Brown and Forsythe tests could be utilized because they are reasonably robust when heterogeneity of variance prevails (Maxwell & Delaney, 1990).

Additionally, some researchers unwisely use omnibus ANOVA tests (i.e., *protected* tests) followed by post-hoc comparisons instead of testing planned contrasts (i.e., *unprotected* tests). Use of omnibus tests when planned comparisons are of interest results in reduced statistical power. We recommend use of planned comparisons as these comparisons tend to be more consistent with carefully structured research questions and serve to limit the number of statistical tests needed (Pedhazur & Schmelkin, 1991). However, in cases in which researchers insist on using omnibus ANOVAs followed by post hoc tests, we suggest that the Dunn-Bonferroni procedure for unprotected tests be utilized because it appears to provide the best control of Type I error (Barnette & McLean, 1998). (Presently, only 12% of researchers use the Dunn-Bonferroni procedure for making pairwise multiple comparisons; Keselman et al., 1998). Moreover, as recommended by Maxwell and Delaney (1990), when conducting pairwise comparisons, the pooled (omnibus) error term should only be used if the variance homogeneity assumption is met--otherwise, a separate error term should be utilized which involves only data representing the levels of interest in the particular comparison. Using the framework of Onwuegbuzie (in press-a), mortality, violated assumptions, and mis-specification error are analytical errors that are particularly pertinent for ANOVA, whereas effect size, confirmation bias, and causal error are pertinent interpretational errors.

Analysis of covariance tests. Most comparisons made in educational research involve

intact groups that may have pre-existing differences. Unfortunately, these differences often threaten the internal validity of the findings (Gay & Airasian, 2000). Thus, in an attempt to minimize this threat, some analysts utilize analysis of covariance (ANCOVA) techniques in which there is an attempt to control statistically for pre-existing differences between the groups being studied.

Prior research suggests that ANCOVA is undertaken in approximately 4% of published research (Elmore & Woehlke, 1988; Goodwin & Goodwin, 1985; Willson, 1980). Unfortunately, most of these published works have inappropriately used ANCOVA because one or more of the assumptions have either not been checked or met--particularly the homogeneity of regression slopes assumption (Glass, Peckham, & Sanders, 1972). As noted by Maxwell and Delaney (1990), ANCOVA represents an ANOVA after adjusting for the regression of the dependent variable on the covariate. In other words, the aim of an ANCOVA is to allocate a percentage of the variance in the dependent variable that would otherwise have been attributed to error in a conventional analysis of variance, to a potentially confounding variable (i.e., the covariate). This partitioning of the variance culminates in a reduction in the sum of squared errors and, consequently, the mean square error. In theory, this subsequent decrease in "noise" helps to clarify the relationship between the independent and dependent variables (Loftin & Madison, 1991).

The first step of ANCOVA is to regress the dependent variable on the covariate, ignoring group membership. After this is undertaken, an ANOVA is conducted on the residualized dependent variable. If the analysis goes as anticipated, holding everything else constant, the ANOVA *F*-statistic is increased because the error variance is smaller than it would have been if the influence of the covariate had not been removed. The all- important assumption that must be met (i.e., homogeneity of regression slopes) implies that the covariate must be highly correlated with the dependent variable but not related to the independent variable. However, as noted by Henson (1998), few covariates exist that meet these criteria--especially when study participants are not randomly assigned to groups (i.e., in quasi-experimental designs), which are endemic to educational research. Unfortunately, if an appreciable correlation exists between the covariate and the independent variable, as is often the case, then the covariate also can *reduce* the variance in the independent variable--culminating in reduced power and effect size. Thus, the homogeneity of regression assumption means that the regression slopes of the covariate and the dependent variable in each group must be identical, or at least similar, if the single pooled regression slope can be utilized accurately with all groups. To the extent that the individual regression slopes are different, the part correlation of the covariate-adjusted dependent variable with the independent variable will more closely mirror a partial correlation, and the pooled regression slope will not provide an adequate representation of some or all of the groups. In this case, the ANCOVA will introduce bias into the data instead of providing a "correction" for the confounding variable (Loftin & Madison, 1991). Ironically, ANCOVA typically is appropriate when used with randomly assigned groups; however, it is typically not justified when groups are not randomly assigned (Henson, 1998).

Another argument against the use of ANCOVA is that after using a covariate to adjust the dependent variable, it is not clear whether the residual scores are interpretable (Thompson, 1992b). Disturbingly, some researchers utilize ANCOVA as a substitute for not incorporating a true experimental design, believing that methodological designs and statistical analyses are synonymous (Henson, 1998; Thompson, 1994a). Thus, we recommend that researchers should use ANCOVA sparingly and with extreme caution. However, when it is utilized, an assessment of the homogeneity of regression assumption always must be undertaken and documented. If the data are shown to violate this assumption (e.g., via a statistically significant Levene test result), the researcher will make a serious mistake in proceeding with the ANCOVA analysis. If ANCOVA is undertaken in the presence of heterogeneity of regression slopes then a Type VII error will prevail (Daniel & Onwuegbuzie, 2000). Using the framework of Onwuegbuzie (in press-a), mortality, violated assumptions, and mis-specification error are analytical errors that are particularly pertinent for ANCOVA, whereas effect size, confirmation bias, and causal error are pertinent interpretational errors.

Multiple regression. In their review of articles published in the *American Education Research Journal, Educational Researcher,* and *Review of Educational Research* over a 20-year period, Elmore and Woehlke (1998) found that multiple regression was the third-most popular statistical technique utilized. Unfortunately, the majority of researchers use multiple regression in inappropriate ways. Undoubtedly the most common error in regression is the use of stepwise regression procedures (i.e., forward selection, backward selection, stepwise selection). Indeed, the use of stepwise regression in educational research is rampant (Huberty, 1994), probably due to its widespread availability on statistical computer

software programs. As a result of this apparent obsession with stepwise regression, as stated by Cliff (1987, pp. 120-121), "a large proportion of the published results using this method probably present conclusions that are not supported by the data."

Perhaps Bruce Thompson has been the most vocal critic of the use of stepwise regression. He and others (Beasley & Leitner, 1994; Davidson, 1988; Edirisooriya, 1995; Lockridge, 1997; Moore, 1996; Thompson, 1994a, 1995, 1998a, 1999; Welge, 1990) have identified at least three problems associated with this technique. First, at every step of the analysis, computer packages use incorrect degrees of freedom in computing statistical significance (Type X error; Daniel & Onwuegbuzie, 2000). Unfortunately, these incorrect degrees of freedom tend to bias statistical significance tests in favor of declaring trivial effects as statistically significant. Second, not only does undertaking *k* steps of analysis not necessarily lead to the best predictor set of size *k*, it is possible that none of the predictors entered in the first *k* steps are even among the best predictor set of size *k*. Third, because the order in which the independent variables are entered in the model is influenced by sampling error, which, at any step, can lead to mis-specification of the model, and because stepwise regression typically involves several steps, this technique often produces results that are very difficult to replicate (Thompson, 1995). A fourth problem identified by the present authors is that because stepwise regression utilizes a series of statistical significance tests, it is subject to actual Type I error rates that can be much greater than its nominal alpha value. For example, a stepwise regression procedure which takes 5 steps to select a final model, with the entry criterion being set at .05 (which is the default value for statistical packages such as SPSS), results in the probability of at least one Type I error rate being .23 (i.e., $1 - (1 - .05)^5$) (see for example, Maxwell & Delaney, 1990). If some variables that are entered are then subsequently removed, then the Type I error rate can increase even more.

Moreover, stepwise regression, more than any other regression technique, tends to capitalize on chance, often resulting in an overfitting of data (Tabachnick & Fidell. 1996) and yielding results that are based on randomness rather than carefully articulated theoretical models. Specifically, decisions about which variables are included in the final regression model are based on *p*-values, which are extremely sample-dependent. For an extensive discussion of the major flaws associated with stepwise regression, see Huberty (1989) and Thompson, Smith, Miller, and Thomson (1991).

As advocated by Thompson (1995), instead of conducting a stepwise regression, an *all possible subsets* (APS) (i.e., *setwise*) multiple regression should be performed. Using this technique, all possible models involving some or all of the independent variables are examined. Indeed, in APS regression, separate regressions are computed for all independent variables singly, all possible pairs of independent variables, all possible trios of independent variables, and so forth, until the best subset of independent variables is identified according to some criterion such as the maximum proportion of variance explained (R^2), which provides an important measure of effect size (Cohen, 1988). Similarly, these repeated subsets can be useful in conducting a regression commonality analysis (Newton & Spurrell, 1967; Rowell, 1996; Seibold & McPhee, 1979). (For an example of an APS multiple regression, see Onwuegbuzie, Slate, Paterson, Watson, and Schwartz, 2000; for an example of commonality analysis, see Daniel, 1989a.) Unfortunately, statistical software programs such as the Statistical Package for the Social Sciences (SPSS; SPSS Inc., 2001) do not allow analysts to conduct APS regression analyses directly, although the Statistical Analysis System (SAS Institute Inc., 1999) does, and no commonly available packages include computations for commonality analyses. It should be noted that both APS regression and commonality analyses represent exploratory model-building tools, as opposed to a model-testing techniques (Tabachnick & Fidell, 1996). As such, APS regression models and commonality analysis results should never be treated as definitive. Rather, they should be subjected to both internal and external replications.

Alternative forms of linear regression techniques that are acceptable are hierarchical (i.e., sequential) multiple regression and standard multiple regression. In hierarchical multiple regression, independent variables are entered into the regression equation in an order specified *a priori* by the researcher. Each independent variable is then evaluated with respect to its own contribution to the model at its own point of entry. Independent variables can be entered one-at-a-time or in blocks in a specified order based on the researcher's theoretical considerations (Tabachnick & Fidell, 1996). Hierarchical regression can be conducted interactively via any statistical software. However, it is likely that many users shy away from this technique because it is not fully automated, unlike stepwise regression. Standard multiple regression involves entering all variables into the regression equation simultaneously (i.e., "direct" variable entry) and assessing the contribution of each (via partial and

semi-partial correlations) as if it had been entered into the model after all other variables had been entered. Both hierarchical multiple regression and standard multiple regression represent model-testing approaches, and are thus recommended for confirmatory purposes.

In any case, whichever technique is used (i.e., APS regression, hierarchical regression, standard regression), it should be noted that the choice of regression variables is just as important as with any other regression technique. That is, the variables that are selected for the initial multiple regression model must be based on theoretical and/or practical considerations (Daniel & Onwuegbuzie, 2001).

In reporting the results of a multiple regression model, most researchers only present unstandardized and/or standardized regression coefficients (b or β weights), and regression structure coefficients (Thompson & Borrello, 1985; Henson, 2002) typically are omitted from the analysis. Yet, structure coefficients, which describe the relationship between scores on a given manifest (i.e., observed) variable with the scores on a given latent (i.e., synthetic) variable, when considered alongside standardized weights, can provide vital information about the relative importance of each of the regression variables (Courville & Thompson, 2001; Henson, 2002). Specifically, the extent to which the standardized weights and the structure coefficients are identical for each variable indicates how uncorrelated the predictor variables are (Thompson, 1998a). Second, if both standardized and structure coefficients of a variable are trivial (i.e., near-zero), the variable is not a practicable predictor of the outcome measure. Third, if a variable has a near-zero standardized coefficient but a large structure coefficient, the variable plays a role in explaining the dependent variable, but the variable is collinear with at least one additional predictor variable. Finally, if a variable has a near-zero structure coefficient but a large standardized coefficient, this indicates that the variable is a suppressor variable. Suppressor variables are variables which assist in the prediction of dependent variables (i.e., they increase the effect size) due to their correlation with other independent variables (Tabachnick & Fidell, 1996). Specifically, suppressor variables improve the predictive power of the other independent variables in the model by suppressing variance that is irrelevant to this prediction, as a result of the suppressor variable's relationship with the other independent variables.

Although the proportion of variance explained (i.e., R^2) is routinely reported, very rarely is the corrected estimate of explained variance (adjusted R^2) reported. Yet, this adjusted measure helps to reduce the positive bias that is inherent in R^2 (Ezekiel, 1930; Wherry, 1931) when sample size is small, correlation is trivial, or the number of predictor variables is large. Another mistake that analysts make when conducting multiple regression analyses is that they do not consider in result interpretation the context-specificity of analytical weights (Thompson, 1998a). This often leads to over-interpretation of regression weights (Cliff, 1987). Thus, as recommended by Thompson (1998a), regression weights always should be interpreted with extreme caution. Additionally, few researchers provide an analysis of the residuals to assess the extent to which the selected multiple regression model fits the underlying data (i.e., meets the regression assumption of constancy of error variance). Even less frequent is the examination of influence diagnostics to determine whether any observations (i.e., cases) exert an undue amount of influence on the regression results (Fox, 1997). Such influence typically is counterproductive; however, if the case diagnostics indicate a serious deficiency in the regression model, it is very likely that the given observation will provide valuable information to the analyst.

As noted by Myers (1986), influence diagnostics include the following: (a) the number of estimated standard errors (for each regression coefficient) that the coefficient changes if the *i*th observation were set aside (i.e., DFBETAS); (b) the number of estimated standard errors that the predicted value changes if the *i*th point is removed from the data set (i.e., DFFITS); and (c) the reduction in the estimated generalized variance of the coefficient over what would have been produced without the *i*th data point (i.e., COVRATIO). (For an example of the use of influence diagnostics, see Onwuegbuzie et al., 2000.)

Most analysts do not appear to evaluate multicollinearity among the regression variables. Multicollinearity leads to inflated regression coefficients or "bouncing betas." Thus, multicollinearity should routinely be assessed in multiple regression models. Techniques for assessing multicollinearity include (a) *variance inflation factors* (VIFs), which indicate the extent to which the variance of an individual regression coefficient has been inflated by the presence of collinearity, and (b) *condition numbers*, which represent the ratio of the largest to the smallest eigenvalues based on a principal components analysis of the regression coefficients yielded by a given analysis, and which serve as measures of the strength of linear dependency among the regression variables (Sen & Srivastava, 1990). VIFs and condition numbers less

than 10 indicate that multicollinearity is not appreciably present (Fox, 1997; Myers, 1986).

Another error that appears to be a common feature of multiple regression analyses is an inadequate case-to-independent variable ratio. Green (1991) recommended using the following guideline for determining an appropriate sample size for a multiple regression analysis that takes into account the effect size. According to Green, the sample size should be greater than or equal to $(8 / F^2) + (I - 1)$, where $F^2 = R^2/(1 - R^2)$. The sample size should exceed this value if the dependent variable is skewed, if one or more of the variables yield low score reliability, or if cross-validation is needed to test the generalizability of the regression model.

Finally, we recommend the use of internal replications, in order to avoid Type V error (internal replication error; Daniel & Onwuegbuzie, 2000). The three most common classes of internal replication utilize either cross-validation, jackknife, or bootstrap techniques (Thompson, 1994b). For regression analyses, cross-validation involves dividing the sample into two approximately equally sized sub-samples (although equality of sub-samples is not required), computing the regression coefficients for the first sub-sample, and then using the second sub-sample to attempt to confirm the results of the first sub-sample. Also, the results of the second subgroup can be confirmed via the first subsample's data. Jackknife techniques involve conducting separate analyses, with groups of participants of an equal size (usually one at a time) being deleted from each analysis once only until all cases/groups have been dropped. The regression results at each stage would be compared to determine stability. Finally, bootstrap methods involve resampling the same dataset repeatedly (i.e., thousands of times), and then computing the regression coefficients and R^2 values for each sample. These coefficients are then compared to the original regression coefficients from the full sample in order to assess stability. Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias, violated assumptions, multicollinearity, and mis-specification error are analytical errors that are particularly pertinent for multiple regression, whereas effect size, confirmation bias, illusory correlation, positive manifold, and causal error are pertinent interpretational errors.

Multivariate analysis of variance/ covariance. With the increased availability of comprehensive statistical software, more researchers are utilizing multivariate statistical techniques. For example, in a review of 36 research articles published in the 1998 volume of the British Journal of Educational Psychology, Onwuegbuzie (2002b) found that multivariate analysis of variance (MANOVA) was the second most common technique utilized. Specifically, this technique was undertaken in nearly one-fourth of the articles examined. Elmore and Woehlke (1998) found that MANOVA was utilized in 12.4% of the articles contained in journals published by the American Educational Research Association from 1978 to 1997.

Unfortunately, several flaws are associated with use of multivariate analyses of variance. For example, some researchers undertake one-way repeated measures analyses of variance (ANOVAs) in order to determine whether there are statistically significant differences among multiple measures (i.e., an omnibus test), and then, if a statistical significant difference is found, follow up with a series of univariate analyses with Type I error rate protection (e.g., Scheffé tests). However, this practice is now outdated. Moreover, many statisticians criticize this technique because analyses involving repeated measures test "linear combinations of the outcome variables (determined by the variable intercorrelations) and therefore do not yield results that are in any way comparable with a collection of separate univariate tests" (Keselman et al., 1998, p. 361). In fact, using ANOVA as a follow-up to MANOVA is a variant of Type IV error (see Daniel & Onwuegbuze, 2000).

Moreover, although as many as 37.5% of researchers conduct a MANOVA followed by a univariate analyses (i.e., a MANOVA-univariate data analysis strategy) (Onwuegbuzie, 2002b), as noted by Keselman et al. (1998, p. 361), "there is very limited empirical support for this strategy. Indeed, Keselman et al. (1998) stated that "If the univariate effects are those of interest, then it is suggested that the researcher go directly to the univariate analyses and bypass MANOVA. . . . Focusing on results of multiple univariate analyses preceded by a MANOVA is no more logical than conducting an omnibus ANOVA but focusing on the results of group contrast analyses (Olejnik & Huberty, 1993)" (pp. 361-362). Furthermore, because this technique relies on a statistically significant MANOVA omnibus test as a precursor to using ANOVA on a *post hoc* basis, the incompatibility of MANOVA and ANOVA, due to the differences in their respective mean square errors and error degrees of freedom, results in a *post hoc* ANOVA test that has lower statistical power than if the ANOVA test had been used as a planned comparison.

Thompson (1999) also criticized researchers who perform several univariate analyses to analyze multivariate data. He maintained that because univariate analyses can be viewed as assessing the contribution of one or more independent variables to a solitary dependent variable, it typically does not

honor, in the optimal sense, the nature of reality that most researchers are interested in studying. This is because most phenomena involve multiple effects. As Tatsuoka (1973) asserted:

> The often-heard argument, "I'm more interested in seeing how each variable, in its own right, affects the outcome" overlooks the fact that any variable taken in isolation may affect the criterion differently from the way it will act in the company of other variables. It also overlooks the fact that multivariate analysis-precisely by considering all the variables simultaneously--can throw light on how each one contributes to the relation. (p. 273)

Thus, we recommend that researchers avoid using the MANOVA-ANOVA analytical strategy, and focus instead on conducting analyses that most appropriately reflect the underlying multivariate reality of interest. (For a more extensive discussion of MANOVA versus multiple ANOVAs, see Huberty and Morris, 1989.)

Also, we suggest that researchers use the multivariate approach to analyzing repeated-measures data (which basis its analysis on the difference scores) rather than the mixed-methods (i.e., with one factor representing the between-subjects factor(s) and the other factor representing the within-subject factor(s)) approach because the latter necessitates an assumption that is not required by the former. Specifically, the mixed-model approach requires a *homogeneity of treatment-difference variances* (i.e., *sphericity*) assumption. Simply put, this assumption requires that every measure must have the same variance, and all correlations between any pair of measures must be the same (Maxwell & Delaney, 1990). However, it should also be noted that the multivariate approach itself requires multivariate normality. As such, researchers always should assess the viability of this assumption.

Another oversight of researchers employing MANOVA techniques is the failure to report the criteria used for determining statistical significance. These criteria include Wilk's Lambda, Pillai's criteria, Hotelling's trace criterion, and Roy's *GCR* criterion. Under certain conditions (e.g., when the independent variable has two levels), the first three criteria are identical. However, there are times when these techniques will yield different *p*-values. Thus, researchers always should specify which criteria were used.

Finally, as for the case of ANCOVA, multivariate analysis of covariance (MANCOVA) should be used with extreme caution. This is because MANCOVA is subject to the same assumptions as for ANCOVA. However, not only is MANCOVA based on the multivariate normal distribution, but it is also assumed that the regression between covariates and the dependent variables in one group is the same as the regression in other groups (i.e., homogeneity of regression) such that using the mean regression to adjust for covariates in all groups is appropriate (Tabachnick & Fidell, 1996). Using the framework of Onwuegbuzie (in press-a), mortality, violated assumptions, and mis-specification error are analytical errors that are particularly pertinent for MANOVA and MANCOVA, whereas effect size, confirmation bias, and causal error are pertinent interpretational errors.

Descriptive discriminant analysis/predictive discriminant analysis. Huberty and his colleagues (Huberty, 1994; Huberty & Barton, 1989; Huberty & Wisenbaker, 1992) have eloquently differentiated between descriptive discriminant analysis (DDA) and predictive discriminant analysis (PDA). According to Huberty (1994), DDA describes the differences on dependent variables that are measured on the interval or ratio scale with respect to a nominally-scaled variable, namely group membership. On the other hand, PDA involves predicting group membership from response variables that are interval- or ratio-scaled. In PDA, the percentage of correct classification is of particular interest, whereas in DDA, the function and structure coefficients are the focus, with the hit rate being immaterial (Thompson, 1998a). Also, as Thompson (1998a) noted, whereas DDA is a member of the general linear model, PDA is not a direct family member. One of the biggest flaws in interpreting DDA results is a failure to interpret both the discriminant function coefficients and the structure coefficients.

Whether DDA or PDA is utilized, many researchers do not report the criteria used for statistical significance (e.g., Wilks' Lambda, Pillai's criteria, Hotelling's trace criterion, and Roy's *GCR* criterion). In addition, many analysts utilize stepwise discriminant analysis techniques. As is the case for stepwise multiple regression, stepwise discriminant analysis contains serious flaws (e.g., Type X error; Daniel & Onwuegbuzie, 2000). Thus, this technique should never be used. Instead, standard discriminant analysis or hierarchical discriminant analysis could be utilized. Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias, violated assumptions, and mis-specification error are analytical errors that are particularly pertinent for discriminant analysis, whereas effect size, confirmation bias, positive manifold, and causal error are pertinent interpretational errors.

Both PDA and DDA are subject to the assumption of multivariate normality. This assumption means that scores on the predictor

variables are independently and randomly sampled from a population, and that the sampling distribution of any linear combination of predictors is normally distributed. Unfortunately, these procedures are not robust to departures from normality if the group sizes are very unequal. Indeed, logistic regression is more appropriate than is discriminant analysis in the presence of non-normality and unequal group sizes (Tabachnick & Fidell, 1996), and thus could be utilized in this case. In fact, logistic regression is more versatile than is discriminant analysis because less stringent assumptions are needed. Specifically, logistic regression makes no assumptions about the distributional properties of the regression variables--in particular, the predictors do not have to be normally distributed; nor do they have to linearly related or have equal variances within each group. Also, the regression variables can be discrete, continuous, or a combination of the two. It is thus surprising how infrequent logistic regression is used in educational research--despite its popularity in the health sciences. Because logistic regression is a discrete response-variable analog to multiple regression, the recommendations made above for the latter (e.g., non-use of stepwise methods, examining residuals, and conducting internal replications) are pertinent for using the former.

Canonical correlation analyses. Canonical correlation analysis is utilized to examine the relationship between two sets of variables when each set contains more than one variable (Cliff & Krus, 1976; Darlington, Weinberg, & Walberg, 1973; Thompson, 1980, 1984, 1991). Indeed, as noted by Knapp (1978, p. 410), "virtually all of the commonly encountered tests of significance can be treated as special cases of canonical correlation analysis." That is, canonical correlation analysis can be used to undertake all the parametric tests which canonical correlation methods subsume as special cases, including Pearson correlation, *t*-tests, multiple regression, analysis of variance, and analysis of covariance (Henson, 2000; Roberts & Henson, 2002; Thompson, 1988b, 1998a, 1991).

Humphries-Wadsworth (1997) reviewed articles published between 1988 and 1998 in which canonical correlation analyses were undertaken. She identified several problems arising from the use of this technique. These problems included inconsistencies in the terminology used to label the same procedure (e.g., "canonical loadings," "canonical weights," "correlation loadings," and "canonical correlates"), and failure to report all the necessary information.

Summarizing Thompson's (1992a) recommendations, Humphries-Wadsworth (1997) stated that when performing a canonical correlation analysis, (a) both the *p*-values pertaining to canonical functions and the squared canonical correlation coefficients (i.e., effect sizes) should be assessed; (b) both the canonical function coefficients and the canonical structure coefficients should be interpreted, along the lines outlined above for multiple regression; (c) redundancy coefficients, which are equal to the average of the squared multiple correlation of each of the variables in one set with all the variables in the other set (Pedhazur, 1982) should not be interpreted because they represent univariate statistics; (d) communality coefficients should be routinely examined; and (e) internal replications (e.g., cross-validation, jackknife, or bootstrap techniques) should be undertaken. Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias, violated assumptions, multicollinearity, and mis-specification error are analytical errors that are particularly pertinent for canonical correlation analyses, whereas effect size, confirmation bias, illusory correlation, crud factor, positive manifold, and causal error are pertinent interpretational errors.

Principal component analysis and factor analysis. Principal component analysis (PCA) and factor analysis (FA) are statistical procedures performed on a set of variables in order to determine which variables in the set form logical subsets that are statistically independent from each other. Specifically, variables that are statistically related with each other but statistically independent from other subsets of variables are combined into components/factors. These components or factors thus are assumed to represent the underlying phenomena/constructs that are responsible for the observed correlations among the variables.

The overall goals of both PCA and FA, which are the two most common methods of factor extraction, are to reduce the dimensionality of the set of variables, to summarize patterns of correlations among manifest variables, to describe an underlying process via the observed relationships among variables, or to test theories about the nature of underlying processes or constructs (Henson et al., 2001; Henson & Roberts, in press; Tabachnick & Fidell, 1996).

There are two major types of factor analysis: exploratory and confirmatory. Exploratory factor analysis (EFA) is an analytic technique conducted in the early stages of the research process with the goal of reducing a larger set of variables into a smaller, interpretable set based on the correlations among the variables. In so doing, the analyst hopes to understand better the internal structure of an instrument or a dataset when insufficient information is available about the data structure. Simply put, exploratory factor analyses are based on

mathematical solutions and do not incorporate *a priori* theoretical underpinnings (Daniel, 1989b). On the other hand, confirmatory factor analysis (CFA) is typically utilized in the latter stages of the research process to test a theory about the latent processes (Henson et al., 2001; Henson & Roberts, in press; Kieffer, 1999).

A common flaw that is apparent in factor analysis is the use of inadequate case-to-variable ratio (Henson et al., 2001; Henson & Roberts, in press). For example, in a review of 40 articles using exploratory factor analysis published in the American Educational Research Journal (Vol. 33-36), Journal of Educational Research (Vol. 89-93), or The Elementary School Journal (Vol. 96-100), Henson et al. (2001) found that 14% of EFAs used subject-to-variable ratios of less than 5:1, with two studies using fewer participants than variables. Similarly, Henson and Roberts (in press) reported that 11.86% of the 60 EFA articles they examined had ratios less than 5:1. Comrey and Lee (1992) suggest that for factor analyses, sample sizes of 50 are very poor, 100 are poor, 200 are fair, 300 are good, 500 are very good, and 1,000 are excellent. Tabachnick and Fidell (1996) and Kieffer (1999) recommend that at least 300 cases be used for factor analysis. However, these guidelines are too simplistic because they do not directly take into account the number of variables. We recommend using case-to-variable ratios as a guideline (Henson et al., 2001; Stevens, 1996, 2002). To this end, we suggest using 5 participants per variable as the bare minimum, although at least 10 participants per variable is much more desirable (Gorsuch, 1983). When researchers use case-to-variable ratios that are less than 5, this should be readily acknowledged in the report as posing a threat to internal validity (i.e., the reliability of the variable scores and emergent factor scores).

The difference in PCA and FA is that the former utilizes the total variance of each variable to assess the shared variation among the variables. That is, PCA uses "ones" on the diagonal of the correlation matrix that is factor analyzed. On the other hand, FA utilizes estimates of common variance or reliability on the main diagonal (Henson et al., 2001; Henson & Roberts, in press; Thompson & Daniel, 1996). It is likely that FA better reflects reality better than does PCA because the latter assumes that each variable represents scores that are perfectly reliable (Kieffer, 1999). Regardless, as noted by Thompson and Daniel (1996), heated arguments prevail as to the relative merits of PCA or FA. Some statisticians (e.g., Daniel, 1990; Thompson, 1992c) have asserted that the difference between PCA and FA is trivial. More specifically, Thompson and Daniel (1996) reported that the difference between PCA and other extraction methods reduces as the number of factored variables increases and as scores on the factored variables become more reliable. However, other researchers (e.g., Gorsuch, 1983) have maintained that there is enough discrepancy between the two procedures to justify careful consideration of which technique to utilize. In any case, our position is that researchers should specify which extraction method they have used and provide a rationale for their choice. Analysts may even want to consider examining both PCA and FA results and then selecting the method which provides the most meaningful interpretation.

As noted by Hetzel (1996), a common misunderstanding among novice factor analysts is incorrectly assuming that the eigenvalue for a specific factor after extraction is identical to the trace (summation of squared values in columns of the factor pattern/structure matrix) after the factor solution is rotated. This error in thinking leads to incorrect proportions of variance being reported for factors. As identified by Thompson (1997), another mistake made by some analysts is a failure to interpret both the factor pattern matrix and the factor structure matrix after conducting an oblique rotation (i.e., rotation of the factors in the factor space such that the angle between the factors is different than 90 degrees). The rationale for this is the same as for interpreting both standardized coefficients and structure coefficients in multiple regression and discriminant analysis. Reporting only one of these two matrices provides only partial information (Henson et al., 2001; Henson & Roberts, in press; Thompson, 1997). On the other hand, when varimax rotation (i.e., orthogonal rotation of the factors in the factor space such that all factors are at 90-degree angles to each other) is utilized, the factor pattern matrix and the factor structure matrices are identical.

Perhaps the most common flaw in articles reporting factor analyses is the lack of attention to detail (Henson et al., 2001; Henson & Roberts, in press). Indeed, of the factor-analytic studies in the field of counseling psychology examined by Tinsley and Tinsley (1987), most did not accurately and completely report the results. In a follow-up study by Hetzel (1996), none of the factor-analytic articles reviewed contained all of the necessary information. In a further replication of Tinsley and Tinsley's (1987) seminal work, Kieffer (1999) had very similar conclusions, as did Henson et al. (2001) and Henson and Roberts (in press).

We recommend that exploratory factor analyses include as many of the following pieces of information as possible: initial number of variables, sample size, sample composition, sampling design, means and variances of the items, correlation matrix

(for replication purposes), method of factor extraction, criteria used for selecting the number of factors to be extracted, method of factor rotation, eigenvalues, correlation matrix of the extracted factors, final communality estimates, estimates of reliability, rotated factor pattern matrix, and rotated factor structure matrix (if oblique rotation is utilized) (cf. Hetzel, 1996). (For an example of how to report EFA results in table form for orthogonal rotations and oblique solutions, see Table 4 of Henson & Roberts, in press, and Henson et al., 2001, respectively.) We recognize that many factor analysts are operating under stringent page restrictions. Nevertheless, attempts should be made to provide as much as the above information as possible. Researchers conducting exploratory factor analysis also should provide an explicit justification for each criterion used in the analytical process (Henson & Roberts, in press). Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias and mis-specification error are analytical errors that are particularly pertinent for exploratory factor analyses, whereas confirmation bias, crud factor, and positive manifold are pertinent interpretational errors.

Confirmatory factor analyses. When performing confirmatory factor analyses, some researchers mistakenly analyze the correlation matrix instead of the variance-covariance matrix (Thompson & Daniel, 1996). Using correlation matrices with confirmatory factor analyses is tantamount to utilizing a variance-covariance matrix wherein the manifest variables have been standardized to unit variance (Bollen, 1989), which likely does not reflect reality. As noted by Skehan (1991), the acceptance or rejection of a confirmatory factor model is not only a function of the difference between the model and reality, but it also is a function of the size of the sample. In particular, large samples tend to have a bias toward rejection of models (Skehan, 1991). According to Schumacker and Lomax (1996, p. 125), for sample sizes larger than 200, "the χ^2 test has a tendency to indicate a significant level" and, consequently, to lead to a rejection of the underlying model. Thus, it is even more important that effect sizes are reported alongside χ^2 values. Indeed, because there does not appear to be a universally agreed-upon index for assessing model adequacy, we recommend that researchers report several fit indices (i.e., effect size measures) such as the ratio of chi-square to degrees of freedom (χ^2/df), the Adjusted Goodness-of-Fit Index, the relative fit index (RFI), the incremental fit index (IFI), the Tucker-Lewis index (TLI), and the comparative fit index (CFI) (Bentler, 1990; Bentler & Bonett, 1980; Bollen, 1986, 1989; Schumacker & Lomax, 1996). Cut-off values between .90 (e.g., Bentler & Bonett, 1980) and .95 (Hu & Bentler, 1999) have been recommended for demonstrating model adequacy.

The root mean square error of approximation (RMSEA; Browne & Cudeck, 1993) is another index that researchers should consider reporting. The RMSEA, which is the square root of the difference between the population covariance matrix and the fitted matrix divided by the number of degrees of freedom for testing the model (i.e., the discrepancy per degree of freedom for the model), is used to compare the fit of two different models to the same data. The RMSEA is bounded below by zero and will be zero only if the model fits exactly (Browne & Cudeck, 1993). Browne and Cudeck (1993) asserted that (1) a RMSEA of approximately .05 or less is indicative of a close fit of the model in relation to the degrees of freedom, (2) a RMSEA value between .05 and .08 indicates a reasonable error of approximation, and (3) models with RMSEA's greater than 0.1 always should be rejected. Hu and Bentler (1999) suggest a cut-off value of .06 for the RMSEA. With respect to the (χ^2/df) ratio, although some researchers (e.g., Carmines & McIver as cited in Arbuckle, 1997) recommend a range between 2 to 1 and 3 to 1 for declaring an acceptable fit, most researchers (e.g., Byrne, 1989) believe that relative chi-square ratios above 2.00 represent an inadequate fit. Thus, we recommend this latter value.

It should be noted, however, that several Monte Carlo studies (i.e., studies in which a series of specific empirical sampling distributions for each index are examined) have demonstrated that many effect size indices also are affected by sample size. For example, Marsh, Balla, and McDonald (1988), who analyzed the distributions of 29 different indices (e.g., GFI, NFI, TLI), found several of these indices to be related to sample size. Notwithstanding, in most cases, all the fit indices obtained using ML techniques tend to perform much better with respect to accuracy of estimates and correctness of statistical results than those obtained using other techniques such as generalized least squares and the asymptotic distribution free method (Hu & Bentler, 1995).

Apart from sample bias, violation of assumptions underlying estimation methods--specifically, violation of distributional assumptions and the effect of dependence of latent variates--can threaten the adequacy of fit indices. In particular, Hu and Bentler (1995) reported that, when latent variables are dependent, most fit indices over-reject models at a sample size of 250 or less. Unfortunately, given that chi-square tests have a tendency to reject models using sample sizes greater than 200, and that most fit indices lead to an over-rejection of models

for samples smaller than 250 when latent variables are dependent, it is difficult, if not impossible, to recommend an ideal sample size for CFA studies. (For comprehensive examples of exploratory factor analysis and confirmatory factor analysis see Kieffer, 1999, and Onwuegbuzie, Bailey, and Daley, 2000.) As for exploratory factor analyses, non-interaction seeking bias and mis-specification error are analytical errors that are particularly pertinent for confirmatory factor analysis, whereas effect size, confirmation bias, crud factor, and positive manifold are pertinent interpretational errors.

Path analysis. Path analysis, which was developed in the 1920s by Sewall Wright in order to gain a better understanding of genetic theory, became popularized in behavioral and social sciences in 1960s (Schumacker & Lomax, 1996). Path analysis is a technique for studying the direct and indirect effects of variables on one or more outcomes. Direct effects involve two variables (observed or latent) that are connected by a single directional path, which represents the regression of the outcome on the predictor. By contrast, indirect effects occur between two latent variables when no single direct path connects them, but instead when the second variable is logically related to the first latent variable through one or more other latent variables via their paths. Conveniently, path coefficients in path models take on the values of Pearson product-moment correlation coefficients or standardized partial regression coefficients. Moreover, the paths suggest whether the dependent variables are related to correlated *effects*, mediated *effects*, and/or independent *effects*. Unlike multiple regression analyses, path analysis models allow analysts to specify the type of relationship among the independent variables when predicting one or more dependent variables.

Path analysis involves decomposing correlations and then comparing original coefficients with the path coefficients computed on the basis of the path model. Correlations between any two variables are decomposed into simple and complex paths (Schumacker & Lomax, 1996). Path coefficients can be tested for statistical significance (e.g., using *t*-values), whereas the overall path model can be tested for goodness of fit using various test statistics (e.g., chi-square tests). Unfortunately, because statistically significant chi-square values suggest that a model does not fit the underlying data, sample sizes greater than 200 have a tendency to reject models, as is the case for CFA. Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias, violated assumptions, multicollinearity, and mis-specification error are analytical errors that are particularly pertinent for path analyses, whereas effect size, confirmation bias, illusory correlation, crud factor, positive manifold, and causal error are pertinent interpretational errors.

Structural Equation Modeling. Structural equation models differ from path analysis in that the former focus on latent variables rather than observed variables, and combine a measurement model (i.e., confirmatory factor analysis) with a structural model (i.e., path analysis) to substantiate theory (Schumacker & Lomax, 1996). By first utilizing multiple observed variables in defining a particular latent variable or hypothesized construct (e.g., a factor), measurement error can be estimated, and, as such, measurement properties (i.e., structural-related validity) can be assessed via parameter estimates.

As with confirmatory factor analysis, we recommend that researchers who utilize structural equation techniques report several fit indexes simultaneously (Thompson, 2000), because there is "*no single* statistical test of significance that identifies a correct model given the sample data" [emphasis in original] (Schumacker & Lomax, 1996, p. 120). Also, as recommended by Schumacker and Lomax (1996), we advocate that a bootstrap analysis should be conducted to determine the stability of path coefficients for the selected model. Bootstrapping involves re-sampling the data (with replacement) a specified (large) number of times to generate statistical estimators adjusted for case-by-case bias and to establish standard error bands around these estimators. These sample bootstrap estimates and standard errors are averaged and used to obtain confidence intervals around the average of the bootstrap estimates (i.e., bootstrap estimators). The bootstrap estimators and their corresponding confidence intervals are then used to determine how stable the sample statistic is as an estimator of the population parameter.

However, it should be noted that even though SEM analyses often lead to models that more closely reflect reality, many of these resultant models may still be under-specified because they (a) do not include interaction effects, (b) do not test for non-linear relationships, and/or (c) fail to account for a sufficient number of observables to identify one or more of the latent variables (i.e., under-identification of the model). Another concern surrounding SEM is the fact that this method of analysis also has been termed *causal modeling*. This is regrettable because the term causal modeling appears to give many researchers the impression that SEM is a method of identifying causes--which is not necessarily the case. Indeed, SEM is no less correlational in analytical framework than is any other member of the general linear model. That is, as is the case for all types of statistical analyses regardless of level of complexity, use of SEM can only allow for causal statements to

be made if the research design permits it (i.e., experimental). Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias, violated assumptions, multicollinearity, and mis-specification error are analytical errors that are particularly pertinent for SEM, whereas effect size, confirmation bias, illusory correlation, crud factor, positive manifold, and causal error are pertinent interpretational errors.

Hierarchical (Multilevel) Linear Modeling. Hierarchical Linear Modeling (HLM) is a technique designed to analyze data that are structured hierarchically. Indeed, HLM has been found to be especially relevant to studies of educational settings because students typically are clustered together within classes, classes are clustered together within schools, schools are clustered together within local education authorities or school districts, and so forth (Bryk & Raudenbush, 1992; Goldstein, 1987, 1995; Gray & Wilcox, 1995; Kreft & De Leeuw, 1998). As HLM software has become more readily available, this method of data analysis is increasing in popularity (Onwuegbuzie, 2002b). However, due to its relative complexity, still relatively few researchers use HLM.

In HLM, models contain one or more variables measured at different levels of the hierarchy. Models can have as few as two levels (e.g., students nested within classes), or many more than two. The lowest level measurements are referred to as being at the *micro level*, whereas all higher-level measurements are deemed to be at the *macro level*. Because HLM models are generalizations of multiple regression models (Kreft & De Leeuw, 1998), the same assumptions associated with multiple regression not only prevail when using HLM, but they are more complicated. Moreover, when these assumptions are violated, Type I and Type II errors will be imminent.

As cautioned by Kreft and De Leeuw (1998), HLM should not be used for data exploration. Indeed, such exploration should be undertaken prior to the HLM stage. Additionally, when using HLM, researchers should refrain from testing models that are too complex--that is models that contain many independent variables, measured at all levels of the hierarchy, and/or that include many cross-level interactions (Kreft & De Leeuw, 1998). Such models are to be avoided, not only because they are sensitive to subtle changes in the system and thus contain unstable parameter estimates, but also because such models are much more difficult to interpret, as well as to replicate from one sample to the next. Using the framework of Onwuegbuzie (in press-a), non-interaction seeking bias, violated assumptions, multicollinearity, and mis-specification error are analytical errors that are particularly pertinent for HLM, whereas effect size, confirmation bias, illusory correlation, crud factor, positive manifold, and causal error are pertinent interpretational errors.

Summary

The purpose of the present paper was to identify and to discuss major analytical and interpretational errors that occur regularly in quantitative and qualitative educational research. With respect to qualitative, interpretivist research, the most common errors are failure to provide evidence for judging the credibility (i.e., validity) of the findings, generalizing findings beyond the sample, and failure to estimate and to interpret effect sizes. Typical errors associated with quantitative research include (a) no evidence provided that statistical assumptions were checked; (b) no power/sample size considerations discussed; (c) inappropriate treatment of multivariate data; (d) use of stepwise procedures; (e) failure to report score reliability indices for either previous or present samples; and (f) no control for Type I error rate.

However, perhaps the most prevalent two errors made in quantitative research, appear across all types of quantitative analyses, namely the incorrect interpretation of statistical significance and the related failure to report and to interpret confidence intervals and effect sizes (i.e., variance-accounted for effect sizes or standardized mean differences) (Daniel, 1998a, 1998b; Ernest & McLean, 1998; Knapp, 1998; Levin, 1998; McLean & Ernest, 1998; Nix & Barnette, 1998a, 1998b; Thompson, 1998b, 2002). This error often leads to under-interpretation of associated *p*-values when sample sizes are small and the corresponding effect sizes are large, and an over-interpretation of *p*-values when sample sizes are large and effect sizes are small (e.g., Daniel, 1998a, 1998c). Because of this common confusion between significance in the probabilistic sense (i.e., statistical significance) and significance in the practical sense (i.e., effect size), some researchers (e.g., Daniel, 1998a) have recommended that authors insert the word "statistically" before the word "significant," when interpreting the findings of a null hypothesis statistical test.

Conclusion

A plethora of analytical and interpretational errors prevails in both quantitative and qualitative research. Based on the frequency of many of the errors identified, one has to wonder what percentage of published educational research findings is invalid. In any case, it is clear that extreme caution should be exercised when undertaking quantitative and qualitative analyses, regardless of level of complexity. Indeed, use of sophisticated analytical techniques and computer software is no substitute for

really getting to know the underlying data and carefully checking all *a priori* assumptions.

We are aware that our views and recommendations provided throughout this essay represent only a portion of the larger body of meta-thinking and appraisal in the field of education that has taken place for many decades. In providing what we believe to be current best practices for various data-analytic techniques, we encourage the reader either to endorse our recommendations or to demonstrate errors in our judgments. At the very least, we hope that we have provided a framework for promoting dialogue.

References

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., with Roediger, H. L., Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). The training in statistics, methodology, and measurement in psychology. *American Psychologist, 45*, 721-734.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing* (rev. ed.). Washington: American Educational Research Association.

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

Arbuckle, J. L. (1997). *AMOS Users' Guide Version 3.6*. Chicago, IL: SmallWaters Corporation.

Barnette, J. J., & McLean, J. E. (1998, November). *Protected versus unprotected multiple comparison procedures*. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Barton, A., & Lazarsfeld, P. F. (1955). Some functions of qualitative data analysis in sociological research. *Sociologica, 1*, 321-361.

Beasley, T. M., & Leitner, D. W. (1994, February). *The p-problem with stepwise multiple regression*. Paper presented at the annual meeting of the Eastern Educational Research Association. (ERIC Document Reproduction Service No. ED 367 669)

Becker, H. S. (1970). *Sociological work: Method and substance*. New Brunswick, NJ: Transaction Books.

Becker, H. S., Geer, B., Hughes, E. C., & Strauss, A. L. (1977). *Boys in white: Student culture in medical school*. New Brunswick, NJ: Transaction Books. (Original work published by University of Chicago Press, 1961)

Bem, S. L. (1981). *Bem Sex-Role Inventory: Professional manual.* Palo Alto, CA: Consulting Psychologists Press.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index, *Psychometrika, 51*, 375-377.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, California: Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models. Applications and data analysis methods*. Newbury Park, CA: Sage.

Byrne, B. M. (1989). A primer of LISREL: *Basic applications and programming for confirmatory factor analytic models*. New York: Springer-Verlag.

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

Cliff, N., & Krus, D. J. (1976). Interpretation of canonical analyses: Rotated vs. unrotated solutions. *Psychometrica, 41*, 35-42.

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

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: John Wiley.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hiilsdale, NJ: Erlbaum.

Connolly, P. (1998). 'Dancing to the wrong tune': Ethnography generalization and research on racism in schools. In P. Connolly & B. Troyna (Eds.), *Researching racism in education: Politics, theory, and practice*. Buckingham, UK: Open University Press.

Constas, M. A. (1992). Qualitative analysis as a public event: The documentation of category development procedures. *American Educational Research Journal, 29*, 253-266.

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.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions.* Thousand Oaks, CA: Sage Publications.

Daniel, L. G. (1989a, March). *Commonality analysis with multivariate data sets*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 314 483)

Daniel, L. G. (1989b, November). *Comparisons of exploratory and confirmatory factor analysis*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR. (ERIC Document Reproduction Service No. ED 314 447)

Daniel, L. G. (1990, November). *Common factor analysis or components analysis: An update on an old debate*. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 325 531)

Daniel, L. G. (1997). Kerlinger's research myths: An overview with implications for educational researchers. *Journal of Experimental Education, 65*, 101-112.

Daniel, L. G. (1998a). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for editorial policies of educational journals. *Research in the Schools, 5*, 23-32.

Daniel, L. G. (1998b). The statistical significance controversy is definitely not over: A rejoinder to responses by Thompson, Knapp, and Levin. *Research in the Schools, 5*, 63-65.

Daniel, L. G. (1998c, December). *Use of statistical significance testing in current "general" educational journals: A review of articles with comments for improved practice.* Paper presented at the annual meeting of the Association for the Advancement of Educational Research, Ponte Vedra, FL.

Daniel, L. G., & Onwuegbuzie, A. J. (2000, November). *Toward an extended typology of research errors.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Daniel, L. G., & Onwuegbuzie, A. J. (2001, February). *Multiple regression: A leisurely primer.* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA.

Darlington, R. B., Weinberg, S. L., & Walberg, H. J. (1973). Canonical variate analysis and related techniques. *Review of Educational Research, 42*, 131-143.

Davidson, B. M. (1988, November). *The case against using stepwise regression methods.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 303 507)

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods.* New York: Praeger.

Dey, I. (1993). *Qualitative data analysis: A user-friendly guide for social scientists.* London: Routledge.

Diamond, P. J., & Onwuegbuzie, A. J. (2001). Factors associated with reading achievement and attitudes among elementary school-aged students. *Research in the Schools, 8*, 1-11.

Edirisooriya, G. (1995, November). *Stepwise regression is a problem, not a solution*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS. (ERIC Document Reproduction Service No. ED 393 890)

Elmore, P. B., & Woehlke, P. L. (1988). Statistical methods employed in *American Educational Researcher and Review of Educational Research* from 1978 to 1987. *Educational Researcher, 17*(9), 19-20.

Elmore, P. B., & Woehlke, P. L. (1998, April). *Twenty years of research methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Ely, M., Anzul, M., Friedman, T., Garner, D., Steinmetz, A. C. (1991). *Doing qualitative research: Circles within circles.* New York: Falmer.

Ernest, J. M., & McLean, J. E. (1998). Fight the good fight: A response to Thompson, Knapp, and Levin. *Research in the Schools, 5*, 59-62.

Eisenhart, M. A., & Howe K. R. (1992). Validity in educational research. In M.D. LeCompte, W.L. Millroy, & J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 643-680). San Diego, CA: Academic Press.

Ezekiel, M. (1930). *Methods of correlational analysis.* New York: Wiley.

Fielding, N., & Fielding, J. (1986). *Linking data.* Beverly Hills, CA: Sage.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks, CA: Sage.

Gay, L. R., & Airasian, P. W. (2000). *Educational research: Competencies for analysis and application* (6th ed.). Englewood Cliffs, N.J.: Prentice Hall.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237-288.

Glesne, C., & Peshkin, A. (1992). *Becoming qualitative researchers: An introduction.* White Plains, NY: Longman.

Goetz, J. P., & Lecompte, M. D. (1984). *Ethnography and the qualitative design in educational research.* New York: Academic Press.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.

Goodwin, L. D., & Goodwin, W. L. (1985). Statistical techniques in AERJ articles, 1979-983: The preparation of graduate students to read educational research literature. *Educational Researcher, 14*(2), 5-11.

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

Gray, J., & Wilcox, B. (1995). *Good school, bad school: Evaluating performance and encouraging improvement.* Buckingham: Open University Press.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499-510.

Greenwald, A.G., Pratkanis, A.R., Leippe, M.R., & Baumgardner, M.H. (1986). Under what conditions does theory obstruct research progress. *Psychological Review, 93,* 216-229.

Hall, B. W., Ward, A. W., & Comer, C. B. (1988). Published educational research: An empirical study of its quality. *Journal of Educational Research, 81,* 182-189.

Halpern, E. S. (1983). *Auditing naturalistic inquiries: The development and application of a model.* Unpublished doctoral dissertation, Indiana University.

Hammersley, M. (1992). Some reflections on ethnography and validity. *Qualitative Studies in Education, 5*(3), 195-203.

Henson, R. K. (1998, November). *ANCOVA with intact groups: Don't do it!* Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Henson, R. K. (2000). Demystifying parametric analyses: Illustrating canonical correlation as the multivariate general linear model. *Multiple Linear Regression* Viewpoints, 26, 11-19.

Henson, R. K. (2002, April). *The logic and interpretation of structure coefficients in multivariate general linear model analyses.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Henson, R. K., & Roberts, J. K. (in press). Exploratory factor analysis reporting practices in published research. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 6). Stamford, CT: JAI Press.

Henson, R. K., Capraro, R. M., & Capraro, M. M. (2001, November). *Reporting practices and use of exploratory factor analyses in educational research journals.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR. (ERIC Document Reproduction Service No. forthcoming)

Hetzel, R. D. (1996). A primer on factor analysis with comments on patterns of practice and reporting. In B. Thompson (Ed.), *Advances in social science methodology* (Vol 4, pp. 175-206). Greenwich, CT: JAI Press.

Hu, L.-T. & Bentler, P. M. (1995). Evaluating model fit. In R.H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage Publications, Inc.

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Huberty, C. J. (1989). Problems with stepwise methods--better alternatives. In B.

Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 43-70). Greenwich, CT: JAI Press.

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

Huberty, C. J., & Barton, R. (1989). An introduction to discriminant analysis. *Measurement and Evaluation in Counseling and Development, 22*, 158-168.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105,* 302-308.

Huberty, C. J., & Wisenbaker, J. (1992). Discriminant analysis: Potential improvements in typical practice. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 2, pp. 169-208). Greenwich, CT: JAI Press.

Humphries-Wadsworth, T. M. (1997, April). *Features of published analyses of canonical results.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Johnson, D. W., & Johnson, F. P. (2000). *Joining together: Group theory and group skills* (7th ed.). Boston, MA: Allyn and Bacon.

Kerlinger, F. N. (1960). The mythology of educational research: The methods approach. *School and Society, 85*, 35-37.

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

Kerlinger, F. N., & Pedhazur, E. (1973). *Multiple regression in behavioral research.* New York: Holt, Rinehart and Winston.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D.,

Keselman, J. C., & Levin, J. R. (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. (1999). An introductory primer on the appropriate use of exploratory and confirmatory factor analysis. *Research in the Schools, 6*(2), 75-92.

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

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

Knapp, T. R. (1998). Comments on the statistical significance testing articles. *Research in the Schools, 5,* 39-42.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage.

Lance, T., & Vacha-Haase, T. (1998, August). *The counseling psychologist: Trends and usages of statistical significance testing.* Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools, 5*, 43-54.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.

Lockridge, J. (1997, January). *Stepwise regression should never be used by researchers.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, Texas. (ERIC Document Reproduction Service No. ED 407 425)

Loftin, L. B., & Madison, S. Q. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 133-147). Greenwich, CT: JAI Press.

Lopez, K. A. (1989, November). *Testing interaction effects without discarding variance.* Paper presented at the annual meeting of the Mid-South Educational Research association, Little Rock, AR. (ERIC Document Reproduction Service No. ED 322 167)

Madill, A., Jordan, A., & Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *The British Journal of Psychology, 91*, 1-20.

Magnusson, D. (1967). *Test theory.* Boston, MA: Addison-Wesley.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62*, 279-299.

Maxwell, J. A. (1996). *Qualitative research design.* Newbury Park, CA: Sage.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Belmont, CA: Wadsworth Publishing Company.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*, 15-22.

McMillan, J. H. (1999). Unit of analysis in field experiments: Some design considerations for educational researchers. (ERIC Document Reproduction Service No. ED 428 135)

McMillan, J. H., Lawson, S., Lewis, K., & Snyder, A. (2002, April). *Reporting effect size: The road less traveled.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Meehl, P. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*, 113-115.

Merriam, S. (1988). *Case study research in education: A qualitative approach.* San Francisco, CA: Jossey-Bass.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Miles, M. B., & Huberman, A. M. (1984). Drawing valid meaning from qualitative data: Toward a shared craft. *Educational Researcher, 13*, 20-30.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Moore, J. D. (1996, January). *Stepwise methods are as bad in discriminant analysis as they are anywhere else.* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 395 041)

Myers, R. H. (1986). *Classical and modern regression with applications.* Boston, MA: Duxbury Press.

Nelson, L. R., & Zaichkowsky, L. D. (1979). A case for using multiple regression instead of ANOVA in educational research. *Journal of Experimental Education, 47*, 324-330.

Newman, I., & Benz, C. R. (1998). *Qualitative-quantitative research methodology: Exploring the interactive continuum.* Carbondale, Illinois: Southern Illinois University Press.

Newton, R. G., & Spurrell, D. J. (1967). Examples of the use of elements for clarifying regression analysis. *Applied Statistics, 16*, 165-176.

Neisser, U. (1998). Rising test scores. In U. Neisser (Ed.), *The rising curve* (pp. 3-22). Washington, DC: American Psychological Association.

Nix, T. W., & Barnette, J. J. (1998a). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*, 3-14.

Nix, T. W., & Barnette, J. J. (1998b). A review of hypothesis testing revisited: Rejoinder to Thompson, Knapp, and Levin. *Research in the Schools, 5*, 55-58.

Olejnik, S., & Huberty, C. J. (1993, April). *Preliminary statistical tests.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Onwuegbuzie, A. J. (2000a, November). *On becoming a bi-researcher: The importance of combining quantitative and qualitative research methodologies.* Symposium presented at the annual meeting of the National Academy of Educational Researchers (NAER), Ponte Vedre, Florida.

Onwuegbuzie, A. J. (2000b, November). *Validity and qualitative research: An oxymoron?* Paper presented at the annual meeting of the Association for the Advancement of Educational Research (AAER), Ponte Vedra, Florida.

Onwuegbuzie, A. J. (2002a). Positivists, post-positivists, post-structuralists, and post-modernists: Why can't we all get along? Towards a framework for unifying research paradigms. *Education, 122*, 518-530.

Onwuegbuzie, A. J. (2002b). Common analytical and interpretational errors in educational research: an analysis of the 1998 volume of the British Journal of Educational Psychology. *Educational Research Quarterly, 26*, 11-22.

Onwuegbuzie, A. J. (in press-a). Expanding the framework of internal and external validity in quantitative research. *Research in the Schools.*

Onwuegbuzie, A. J. (in press-b). Effect sizes in qualitative research: A prolegomenon. *Quality & Quantity: International Journal of Methodology.*

Onwuegbuzie, A. J., Bailey, P., & Daley, C. E. (2000). The validation of three scales measuring anxiety at different stages of the foreign language learning process: The input

anxiety scale, the processing anxiety scale, and the output anxiety scale. *Language Learning, 50*(1), 87-117.

Onwuegbuzie, A. J., & Daniel, L. G. (2002a). Uses and misuses of the correlation coefficient. *Research in the Schools, 9*, 73-90.

Onwuegbuzie, A. J., & Daniel, L. G. (2002b). A framework for reporting and interpreting internal consistency reliability estimates. *Measurement and Evaluation in Counseling and Development, 35*, 89-103.

Onwuegbuzie, A. J., & Daniel, L. G. (in press). Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences. *Research in the Schools.*

Onwuegbuzie, A. J., Slate, J., Paterson, F., Watson, M., & Schwartz, R. (2000). Factors associated with underachievement in educational research courses. *Research in the Schools, 7*(1), 53-65.

Onwuegbuzie, A. J., & Teddlie, C. (2002). A framework for analyzing data in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 351-383). Thousand Oaks, CA: Sage.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

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

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Peet, M. W. (1999, November). *The importance of variance in statistical analysis: Don't throw the baby out of the bathwater.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, Alabama.

Polkinghorne, D. (1983). *Methods for the human sciences.* Albany, New York: University of New York Press.

Prosser, B. (1990, January). *Beware the dangers of discarding variance.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Reproduction Service No. ED 314 496)

QSR International Pty Ltd (2002). *N6 (Non-numerical Unstructured Data Indexing Searching & Theorizing) qualitative data analysis program.* (Version 6.0) [Computer software]. Melbourne, Australia: QSR International Pty Ltd.

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement, 62*, 241-253.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rowell, R. K. (1996). Partitioning predicted variance into constituent parts: How to conduct commonality analysis. In B. Thompson (Ed.), *Advances in social science methodology* (Vol 4, pp. 33-44). Greenwich, CT: JAI Press.

Sandelowski, M. (2001). Real qualitative researchers don't count: The use of numbers in qualitative research. *Research in Nursing & Health, 24*, 230-240.

SAS Institute Inc. (1999). *SAS/STAT User's Guide* (Version 6.12) [Computer software]. Cary, NC: SAS Institute Inc.

Schmidt, F .L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Sechrest, L., & Sidani, S. (1995). Quantitative and qualitative methods: Is there an alternative? *Evaluation and Program Planning, 18*, 77-87.

Seibold, D. R., & McPhee, R. D. (1979). Commonality analysis: A method for decomposing explained variance in multiple regression analysis. *Human Communication Research, 5*, 355-365.

Sen, A. K., & Srivastava, M. (1990). *Regression analysis: Theory, methods, and applications.* New York: Springer-Verlag.

Simmelink, S., & Vacha-Haase, T. (1999). *Reliability generalization with the Rosenberg Self-Esteem Instrument.* Paper presented at the annual meeting of the Rocky Mountain Psychological Association, Fort Collins, CO.

Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition, 13*, 275-298.

Smith, J. K., & Heshusius, L. (1986). Closing down the conversation: The end of the quantitative-qualitative debate among educational inquirers. *Educational Researcher, 15*, 4-13.

Snyder, P. A., & Thompson, B. (1998). Use of tests of statistical significance and other analytical choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly, 13*, 335-348.

SPSS Inc. (2001). *SPSS 11.0 for Windows.* [Computer software]. Chicago, IL: SPSS Inc.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins College Publishers.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches.* Applied Social Research Methods Series (Vol. 46). Thousand Oaks, CA: Sage.

Tatsuoka, M. M. (1973). Multivariate analysis in educational research. In F.N.Kerlinger (Ed.), *Review of Research in Education* (pp. 273-319). Itasca, IL: Peacock.

Thompson, B. (1980, April). *Canonical correlation: Recent extensions for modelling educational processes.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretations.* Newbury Park, CA: Sage Publications. (ERIC Document Reproduction Service No. ED 199 269)

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

Thompson, B. (1988a). Discard variance: A cardinal sin in research. *Measurement and Evaluation in Counseling and Development, 21*, 3-4.

Thompson, B. (1988b, April). *Canonical correlation analysis: An explanation with comments on correct practice.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 295 957)

Thompson, B. (1991). Methods, plainly speaking: A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development, 24*, 80-93.

Thompson, B. (1992a, April). *Interpreting regression results: Beta weights and structure coefficients are both important.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Thompson, B. (1992b, April). Misuse of ANCOVA and related "statistical control" procedures. *Reading Psychology: An International Quarterly, 13*, iii-xvii.

Thompson, B. (1992c). A partial test distribution for cosines among factors across samples. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 2, pp-81-97). Greenwich, CT: JAI Press.

Thompson, B. (1994a). *Common methodological mistakes in dissertations, revisited.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA (ERIC Document Reproduction Service No. ED 368 771)

Thompson, B. (1994b). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality, 62*, 157-176.

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. (1997). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. *Educational and Psychological Measurement, 57*, 5-19.

Thompson, B. (1998a, April). *Five methodological errors in educational research: The pantheon of statistical significance and other faux pas.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Thompson, B. (1998b). Statistical testing and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*, 33-38.

Thompson, B. (1999, April). Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap. Invited address presented at

the annual meeting of the American Educational Research Association, Montreal [On-line]. Available: http://acs.tamu.edu/~bbt6147/aeraad99.htm

Thompson, B. (1999b). Improve research clarity and usefulness with effect size indices as supplements to statistical significance testing. *Exceptional Children, 65*(3), 329-337.

Thompson, B. (2000). 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. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32.

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

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56,* 197-208.

Thompson, B., Smith, Q. W., Miller, L. M., & Thomson, W. A. (1991, January). *Stepwise methods lead to bad interpretations: Better alternatives.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the *Journal of Experimental Education. Journal of Experimental Education, 66,* 75-83.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent JCD research articles. *Journal of Counseling and Development,* 76, 436-441.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174-195.

Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology, 34,* 414-424.

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., 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., & Ness, C. (1999). Statistical significance testing as it relates to practice: Use within *Professional Psychology. Professional Psychology Research and Practice, 30,* 104-105.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients. A review of three journals. *The Journal of Experimental Education, 67,* 335-341.

Vacha-Haase, T., & Nilsson, J. E. (1998). Statistical significance reporting: Current trends and usages with *MECD. Measurement and Evaluation in Counseling and Development, 31,* 46-57.

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*(3), 413-425.

Vockell, E. L., & Asher, W. (1974). Perceptions of document quality and use by educational decision makers and researchers. *American Educational Research Journal, 11,* 249-258.

Ward, A. W., Hall, B. W., & Schramm, C. E. (1975). Evaluation of published educational research: A national survey. *American Educational Research Journal, 12,* 109-128.

Welge, P. (1990, January). *Three reasons why stepwise regression methods should not be used by researchers.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED 316 583)

Wherry, R. J., Sr. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2,* 440-457.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F^* statistics. *Communications in Statistics-Simulation and Computation, 15,* 933-943.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. *Educational Researcher, 9*(6), 5-10.

Witcher, A. E., Onwuegbuzie, A. J., & Minor, L. C. (2001). Characteristics of effective teachers: Perceptions of preservice teachers. *Research in the Schools, 8*, 45-57.

Witta, E. L., & Daniel, L. G. (1998, April). *The reliability and validity of test scores: Are editorial policy changes reflected in journal articles?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wolcott, H. F. (1990). On seeking--and rejecting--validity in qualitative research. In E.W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 121-152). New York: Columbia University, Teachers College Press.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). Measurement and evaluation in the schools. New York: Longman.

2003 Article Citation
Onwuegbuzie, A. J., & Daniel L. G. (2003, February 19). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education* [On-line], *6*(2). Available: http://cie.ed.asu.edu/volume6/number2/

Author Notes

Anthony J. Onwuegbuzie
Howard University
Department of Human Development and Psychoeducational Studies, School of Education
2441 Fourth Street, NW, Washington, DC 20059
TONYO@SURFSOUTH.COM

Anthony J. Onwuegbuzie is an associate professor of Educational Psychology at Howard University in Washington, DC. He earned his Ph.D. in Educational Research and two of his three master's degrees (i.e., M.S. in statistics and M.Ed. in Testing and Measurement) at the University of South Carolina. Also, he earned a postgraduate diploma in statistics at the University College London. His research topics primarily involve disadvantaged and under-served populations such as minorities, learning disabled students, and juvenile delinquents. In the last four years, he has secured more than 120 publications in reputable refereed journals. To date, he has made or has been invited to make more than 200 presentations and keynote addresses at the international, national, regional, and university levels, presenting on the continents of North America (including Canada), Europe, and Africa. He has a forthcoming book entitled: *Library anxiety: Theory, research, and applications* (Scarecrow Press).

Larry G. Daniel
University of North Florida

Larry G. Daniel is Associate Dean, College of Education and Human Services, and Professor, Division of Educational Services and Research, at the University of North Florida. His research and writing interest include research methodology and statistics, measurement issues, educational leadership, and teacher education.

Note from the 2015 Executive Editor, Constantin Schreiber

May 22, 2015. This article was first published at the original *Current Issues in Education* website, located at http://cie.asu.edu/articles/index.html. In 2009, *CIE* changed online platforms to deliver the journal at http://cie.asu.edu. The original *CIE* website was from then on only used as an archival repository for published articles prior to Volume 12. After the new *CIE* website moved to a different server in 2014, the original website and original article URLs could not be accessed anymore. Therefore, this article had to be repurposed into the published format you are viewing now.

All content from the original publication has been preserved. No content edits occurred. Spelling, grammar, and mechanical errors that may be found were present in the original publication. The *CIE* logo and publisher information in use at the time of the article's original publication is unaltered. Please direct questions about this article's repurposing to cie@asu.edu.

2015 Article Citation
Onwuegbuzie, A. J., & Daniel, L. G. (2003). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education*, *6*(2). Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/1609

Current Issues in Education

Mary Lou Fulton College of Education
Arizona State University