ED 449 166	TM 032 244
AUTHOR	Daniel, Larry G.; Onwuegbuzie, Anthony J.
TITLE	Towards an Extended Typology of Research Errors.
PUB DATE	2000-11-00
NOTE	36p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (28th, Bowling Green, KY, November 17-19, 2000).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Classification; *Error Patterns; Research Methodology; *Research Problems; *Statistical Significance
IDENTIFIERS	Type I Errors; Type II Errors; Type III Errors

ABSTRACT

This paper proposes a new typology for understanding common research errors that expands on the four types of error commonly discussed in the research literature. Examples are presented to illustrate Type I and Type II errors, errors related to the interpretation of statistically significant and nonsignificant results respectively, with attention given to the control of both types of error. Then an explanation of typical errors that fall into the Type III (incorrect inferences about result directionality) and Type IV ("effects error") categories is offered, along with examples of erroneous conclusions researchers draw when committing these errors. Six additional types of errors are proposed for addition to the typology, with a discussion of how attention to these newly identified error categories can be useful in improving the quality of educational research. These error types are: (1) internal replication error; (2) reliability generalization error; (3) heterogeneity of variance/heterogeneity of regression error; (4) test statistic distribution error; (5) sampling bias error; and (6) degrees of freedom error. (Contains 66 references.) (SLD)

Running Head: TYPOLOGY OF RESEARCH ERRORS

errors.paper.2000

Towards an Extended Typology of Research Errors

Larry G. Daniel
University of North Florida

Anthony J. Onwuegbuzie
Valdosta State University

Paper presented at the annual meeting of the Mid-South Educational Research Association, November 15-17, 2000, Bowling Green, KY.

Abstract

The present paper proposes a new typology for understanding common research errors which expands upon the four types of error commonly discussed in research literature. Examples are presented to illustrate Type I and Type II errors, with attention given to control of both types of error. Next, an explanation of typical errors that fall into the Type III and Type IV categories is offered, along with examples of erroneous conclusions researchers draw when committing these errors. Finally, 6 additional types of error are proposed for addition to the typology, with discussion of how attention to these newly-identified error categorizations can be useful in improving the quality of educational research.

Towards an Extended Typology of Research Errors

Careful design of research is essential if conclusions stemming or arising from any given study are to be meaningful. Serious flaws in a study's design or major errors in the interpretation of a study′ s findings raise serious reservations as to the generalizability of the research findings. It is often the case that the researcher fails to control for extraneous variables that may affect the results or fails to utilize a research design that is robust enough to rule out threats to the internal and external validity of the study. When these failures occur, the research findings are challenged by "rival hypotheses" that serve as competing explanations of the findings. As noted by Cook and Campbell (1979), rival hypotheses may result from various internal validity threats (e.g., history, maturation, mortality, instrumentation, testing, statistical regression, selection, interactions with selection, compensatory rivalry, resentful demoralization, diffusion of treatment) or external validity threats (e.g., selection-treatment interaction, setting-treatment interaction, history-treatment interaction).

Obviously, a major component of generalizability of results is replication. Replication, which has oft been called the "hallmark of science," allows for maximal confidence in research results:

> replication is an important way to increase the validity of generalizing results to varying populations and settings. Replication refers to the repetition of a research study in a new setting (and often by a different researcher) to see whether similar results will be obtained. To the extent that results are replicated

> with varying populations and different settings with variations in the operational definitions of the research variables, it becomes increasingly likely that a generalization to other populations will be valid. (Vockell & Asher, 1995, p. 344)

Despite the importance of full replication, it is also important that researchers realize the importance of assessing generalization of results within the confines of a single study considering that replication may be timely and/or costly.

In addition to controlling for various validity threats, it is also important that the researcher controls for various sources of error that may contaminate data and thereby affect the interpretation of the results, including (a) use of inadequate or biased samples, (b) inappropriate specification of one or more of the study's variables (Pedhazur & Schmelkin, 1991), (c) failure to consider the effects of the reliability and/or validity of data on the statistical results (Crocker & Algina, 1986; Onwuegbuzie & Daniel, 2000a), (d) inadequate statistical power (Cohen, 1988); (e) violated assumptions of statistical tests (Glass, Peckham, & Sanders, 1972; Keselman et al.,1998), (f) use of multiple parametric statistical tests without controlling for inflation of Type I error (Stevens, 1996), and (g) use of univariate statistical tests when multivariate tests more appropriately reflect the reality of the data in hand (Fish, 1988).

Results may also be contaminated by data interpretation errors (Daniel, 1998, 1999; Hall, Ward, & Comer, 1988; Keselman et al., 1998; Onwuegbuzie, 1999; Onwuegbuzie & Daniel, in press; Thompson, 1998a; Vockell & Asher, 1974; Ward, Hall, & Schramm, 1975; Witta & Daniel, 1998). Some two decades ago, Games (1978, p. 257) observed, "There is no statistical technique that will always lead us to 'truth,' nor any set of procedures that cannot be misused." Similarly, Marascuilo and Levin (1970, p. 397) quipped, "To err is human, and now that behavioral researchers are engaging in inferential activity with increasing frequency, it is more than likely that the number of erroneous inferences is also increasing."

Data interpretation errors result in part from shortcomings in the way in which researchers are trained (Kerlinger, 1960; Newman & Benz, 1998) and from proliferations of various erroneous "mythologies" about the nature of research (Daniel, 1997; Kerlinger, 1960). Chief on the list on interpretation errors are various misunderstandings of results of statistical significance testing. While problems and misunderstandings associated with statistical significance testing have been given considerable attention for many years (e.g., Bakan, 1966; Berkson, 1938, 1942; Chow, 1988; Cohen, 1994; Daniel, 1998; Gold, 1969; McLean & Ernest, 1998; Nix & Barnette, 1998; Rozeboom, 1960; Thompson, 1989, 1996), with some researchers even calling for the abandonment of statistical significance testing (e.g., Carver, 1978, 1993), the statistical significance testing paradigm continues to thrive in studies focused on the testing of hypotheses regarding quantitative data.

The Roman Numeral Typology of Research Errors

As the above discussion illustrates, research errors may result from a myriad of sources. Consequently, researchers have conceptualized various frameworks for understanding errors in an attempt to improve the practice of research. The most well-known framework of this type utilizes a series of successive Roman numerals to refer to the various types of error. Four such types of error (Types I, II, III, and IV) have been codified. Neyman and Pearson are generally credited with codifying the original two error types (Types I and II) in their expansion of Fisher's conceptualization of the null hypothesis test (Gigerenzer et al., 1989). In 1970, Marascuilo and Levin noted that there was some discrepancy as to the definition of the term "Type III error," though most today would utilize the term according to Kaiser′ s (1966) conceptualization as referring to an error in interpreting the directionality of a relationship. Finally, Marascuilo and Levin (1970, 1976; Levin & Marascuilo, 1972) coined the term ‶Type IV error″ to refer to errors relative to interaction effects.

Research training in the social sciences typically gives students at least some introduction to Type I and Type II errors, errors related to interpretation of statistically significant and statistically nonsignificant results, respectively. Type III (incorrect inferences about result directionality) and Type IV (‶effects error") errors, although not mentioned with as much frequency in research training, are occasionally given at least some attention. A wealth of literature has been written proposing various other ways of codifying error types. Nevertheless, it is our opinion that the Roman numeral scheme of identifying error types is appealing due to its simplicity. The use of the system of Roman numerals to label error types serves to simplify labels for errors within a system known for the cumbersome use of potentially confusing jargon. Consequently, we believe that a considerable portion of the wealth of excellent writing on the improvement of educational research can be communicated clearly via an expansion of this typology.

Purpose

The present paper proposes an expansion of the "Roman numeral″ typology for understanding common research errors and focuses on (a) a review of Type I and Type II errors, (b) an explanation of the oft-forgotten Type III and Type IV errors, and (c) the recommendation that six additional types of error be added to this typology. Examples are presented to illustrate Type I and Type II errors, with attention given to control of both types of error. Next, an explanation of typical errors that fall into the Type III (incorrect inferences about result directionality) and Type IV ("effects error") categories is offered, along with examples of erroneous conclusions researchers draw when committing these errors. Finally, building on our earlier work in codifying research errors (Onwuegbuzie & Daniel, in press), six additional types of error are discussed and codified.

A Review of Type I and Type II Errors

Type I and Type II errors result from decisions made in relationship to the testing of null hypotheses. In null hypothesis testing, the researcher proposes a relational hypothesis that he/she attempts to nullify based on the data in hand selected from a population of interest. Although the null hypothesis may be any hypothesis that the researcher attempts to nullify (Cohen, 1988), the most common type of null is the Anil@ null (i.e., a hypothesis of no difference or no relationship). When posing a null hypothesis (H_0), the researcher begins with two assumptions: (a) that the null is true in the population of interest, and (b) that the sample is representative of the population of interest. The researcher also determines an alternative hypothesis that he/she believes will more appropriately reflect the results of the statistical analysis if there is enough evidence to reject the null hypothesis.

Once the hypotheses are determined, the study is conducted, data are collected, and a statistical result is obtained using the sample employed for the study with the intention that the result will generalize to the population. The null hypothesis is then consulted to determine the degree to which it reasonably explains the obtained result. This determination is made based on a predetermined probability level (i.e., the "alpha" or "critical probability level") denoting the likelihood of the obtained result if one assumes that the null hypothesis is true in the population. The researcher typically sets alpha (α or p_{critical}) at a very conservative level (5% or less), with values of 5% (p_{critical} = .05) and 1% (p_{critical} = .01) common in the social sciences. The obtained (calculated) statistical finding (i.e., the "test statistic") is then tested against the critical value for that statistic based on the researcher's selected alpha level and a function of the size of the sample (i.e., degrees of freedom).

For example, suppose a researcher uses the Pearson product-moment correlation coefficient (*r*) to measure the linear relationship between two variables for a sample of 100 persons. The null hypothesis would state that $r = 0$, and a reasonable alternative "difference" hypothesis would be $r \neq 0$. If the obtained *r* value is .30, the researcher could conclude that the result is statistically significant at the .01 alpha level based on a comparison with the critical values of *r* from a table in a statistics textbook. By stating that this result is statistically significant at the 1% alpha level, the researcher would conclude that a correlation between the two variables as large as $|.30|$ is less than 1% likely in the population of interest under the assumption that the null hypothesis is true. Alternately, the researcher could conclude that the 1% represents the *conditional* probability of obtaining a correlation of |.30| or larger given the null hypothesis is true. The researcher would reject the null hypothesis, concluding that the null is not a reasonable hypothesis for the data in hand. The researcher would then consult the study′ s alternative hypothesis and make conclusions about the findings based on this hypothesis.

Type I Error (False Positive Error)

Any time that a researcher obtains a statistically significant result, there is some degree of risk that the null hypothesis has been rejected in error (i.e., that the null hypothesis that has been rejected is actually true in the population of interest). This risk, which is related to the selected p_{critical} level, is known as the probability of a Type I error (or α error). In the foregoing correlation example, the null hypothesis was rejected at the .01 level of statistical significance; hence, the researcher′ s risk of rejecting the null when it is really true in the population of interest was less than 1%. Obviously, if the null is true in the population, the correlation between the two variables would be exactly zero. Because the data in hand yielded an *r* of .30, one would conclude that sample bias had contributed to the results under the assumption of a true null. Hence, Type I error is essentially sampling error.

Because the exact characteristics of the population of interest are rarely known, it is difficult to determine whether a given sample is truly representative of the population. In fact, even if the population=s characteristics are known, and the researcher can document that the sample is identical to the population across a variety of known characteristics, there may still be other characteristics across which the sample and the population differ. For this reason, researchers usually employ relatively conservative levels for alpha. Although utilization of a conservative alpha level is important in the researcher"s protection against Type I error, it is also important that the researcher not risk increasing the probability of Type I error by performing multiple statistical tests with the same data set. This latter problem is known as "familywise" Type I error, "probability pyramiding," or "the error rate per experiment" (Cook & Campbell, 1979, p. 43). Hence, researchers are cautioned to limit their research to those null hypotheses that they find most essential to the purpose of the given study (Stevens, 1996) and/or select statistical tests wisely so as to limit the number of tests performed (e.g., multivariate tests) while ideally also maximizing the opportunity to honor the reality to which the researcher is hoping to generalize (Fish, 1988). In order to accomplish these goals, researchers are admonished to develop their hypotheses from a sound theoretical framework that is based on the extant literature.

Type II Error (False Negative Error)

Type II error (beta error) is closely related to Type I error and statistical significance testing. However, whereas Type I error is probable when statistical significance is obtained, Type II errors may ensue when obtaining a statistically nonsignificant result, thereby supporting the tenability of the null hypothesis. Hence, Type II error is the likelihood of failing to reject a null hypothesis that is false in the population of interest. For example, in an experimental study comparing the effect of two teaching methods, A and B, on reading achievement, the researcher might utilize a *t*-test to compare the achievement of the two groups studied on a final reading achievement measure. If in the population of interest, method A is more effective than method B, but in the sample utilized in the study, the achievement level is found to be the same for the A and B groups (due to sampling bias, imprecision of the achievement measure, misspecification of the dependent variable, etc.), the researcher would fail to reject the null hypothesis in error.

Just as the probability of Type I error can be set via the researcher′ s preselection of a value for alpha, the probability of Type II error may also be preselected prior to a study via a statistic known as beta (β); however, it is somewhat more difficult to preselect beta than alpha. Obviously, because Type I and Type II errors result from opposite statistical decisions, attempts to minimize Type I error (by setting a more conservative alpha level) actually serve to increase the likelihood of Type II error, and vice-versa. "Statistical power" (or, simply, "power") is a commonly-used term to indicate the degree to which a result is not affected by a Type II error. Statistical power is mathematically determined by subtracting beta from one ($1 - \beta$); hence, the use of a "high powered" test will increase the likelihood that a given statistical test will detect the presence of a statistical effect (e.g., a group difference) when there actually is a difference in the population of interest, thereby reducing the likelihood of a Type II error. In an attempt to explicate the concept of power in everyday terms, Hays (1988, p. 266) observed,

> The power of a test of H_0 is not unlike the power of a microscope. It reflects the ability of a decision rule [alpha] to detect from evidence that the true situation differs from a hypothetical one. Just as a high-powered

> microscope lets us distinguish gaps in an apparently solid material that we would miss with low power or the naked eye, so does a high-powered test of H_0 almost ensure us of detecting when H_0 is false.

In a group comparison study, statistical power tends to increase when sample size is larger, when the probability of rejecting a null (i.e., the value of α) is increased, when scores are highly reliable (Onwuegbuzie & Daniel, 2000a) or when the statistic to be compared across groups (usually the mean) differs to a larger degree. This difference in means is commonly referred to as the "statistical effect size." Researchers who have some predetermined idea as to what constitutes a reasonable effect size may use this information along with the alpha level they desire in determining optimal sample sizes for a given study in order to increase statistical power (Cohen, 1988). This process is commonly known as "power analysis." Researchers, using power analysis, as well as other procedures to assure an appropriate research design (Benton, 1991), can do much to minimize the likelihood of Type II error. Unfortunately, many researchers ignore the fact that statistical significance of an effect is largely dependent on sample size Onwuegbuzie (2000a), with the typical level of power for medium effect sizes in the behavioral and social sciences disturbingly hovering around .50 (Cohen, 1962). Thus, it is desirable that more researchers conduct *a priori* power analyses.

Explanation of Type III and Type IV Errors

Researchers tend to spend so much time on "the twin gremlins, Type I and Type II errors" (Marascuilo & Levin, 1970, p. 397) that they sometimes forget to consider the potential for Type III and IV errors. Nevertheless, errors of these latter two types are costly to result interpretation, and considering the frequency of these errors (Harwell, 1998; Kaiser, 1966; Levin & Marascuilo, 1972; Marascuilo & Levin, 1970, 1976; Umesh, Peterson, McCann-Nelson, & Vaidyanathan, 1996), it is important that researchers engaged in null hypothesis testing gain an understanding of Type III and IV errors so as to avoid them in their own research.

Type III Error (Directionality Error)

Type III error is the likelihood of making an error in the directionality of a statistical result. For example, in a correlational study, the researcher might find a positive correlation between variables when, in reality (i.e., in the population of interest), the variables are negatively correlated. Similarly, in a group comparison study, the researcher might find that Group A outperforms Group B; yet, in the population, treatment B might actually be more effective than treatment A. This sort of erroneous conclusion is what Huck (2000) has referred to as "the worst kind of inferential error" (p. 197) because it can potentially lead the researcher to consider a view of relationships among variables that is contrary to reality.

There is no set formula for tracking Type III error, but researchers would be wise to employ some sort of sample splitting and cross-validation whenever possible (Oxford & Daniel, in press) so that internal replication may be employed to gather a first estimate of result generalizability and also as a check against Type III error. More sophisticated sample splitting procedures (i.e., jackknife and bootstrap procedures) can also be employed to further examine sample bias and to offer initial evidence of result generalizability (Daniel, 2000). Additionally, when some reasonable number of studies exist in a given area, meta-analysis (Glass, McGaw, & Smith, 1981) can be used as a check against Type III error.

Type IV Error (Effects Error)

Type IV error (Marascuilo & Levin, 1970) occurs when a researcher offers an incorrect follow-up interpretation to a correctly rejected statistical hypothesis. For example, a researcher might follow "a rejected analysis of variance (ANOVA) hypothesis with a set of overlapping multiple *t* tests, each performed at the same alpha level as chosen for the original *F* test" (Levin & Marascuilo, 1972, p. 368). Unfortunately, this procedure may yield statistically significant *t* values that would not have emerged had the researcher used pairwise comparisons (e.g., Scheffé tests) that more appropriately controlled for inflation of Type I error. These results would constitute a Type IV error. These errors may also be found in some abundance in multivariate cases. For example, researchers who follow up statistically significant MANOVAs with multiple univariate ANOVAs increase the likelihood of arriving at Astatistical decisions that are different from those based on multivariate post hoc techniques@ (Levin & Marascuilo, 1972, p. 370).

Type IV errors are particularly problematic in the testing of interaction effects. Researchers who find statistically significant interaction effects in ANOVA, for example, may incorrectly (due to lack of forethought or ignorance of the nature of interaction effects) follow these results up with pairwise or "nested" comparisons of various cell means. To avoid errors of this type, researchers should consider planned contrasts for decomposing complex interaction effects or else more carefully designed post hoc procedures focusing on cell means that better honor the original interaction hypotheses (Dodds, 1998). Careful selection of the tests used to break down complex interaction effects in these models will serve to eliminate Type IV errors from research. These strategies will also serve to minimize experimentwise Type I error rates.

Expanding the Roman Numeral Typology of Errors

Anyone spending any appreciable amount of time in the literature on research quality has noted the numerous calls for quelling various errors related to the use and interpretation of research and statistical procedures. In an effort to further codify a number of the problems we have seen identified in this literature, we propose that six additional error types be added to the four types presently defined via the Roman numeral typology. Our goal herein is not necessarily to add new insights regarding the problems we address, but rather to offer an organizational schema whereby these errors can be more commonly recognized and avoided by social science researchers. Our expanded typology is illustrated in Table 1. The narrative to follow offers commentary on the six additional error types included in the table.

INSERT TABLE 1 ABOUT HERE

Type V Error (Internal Replication Error)

Although replication within the social sciences is of paramount importance, and although repetition of studies with different samples in different settings (external replication) is essential to advancement of a field of study, many researchers do not realize that first estimates of replicability of results can be gathered within the confines of individual studies (internal replication). For example, except when sample size is extremely small, it is advantageous for researchers to split samples in half and conduct their statistical analyses with both subsamples followed by "cross-validation" of the findings (Daniel, 2000). As illustrated by Oxford and Daniel (in press), cross-validation involves taking statistical weights derived from one subsample's data and using these weights to develop statistical estimates for the other subsample. As an alternative to basic cross-validation, researchers may also used the more sophisticated "data intensive" sample reconstitution procedures, including the "bootstrap" (repeated resamplings with replacement from a single sample followed by a repetition of the analysis after each resampling) and the "jackknife" (repetition of an analysis *n* times using a sample of size *n,* with one unique observation removed at each repetition followed by computation of a weighted estimate of the statistic based on the *n* repetitions).

As Onwuegbuzie and Daniel (in press) indicated, internal replications featuring jackknife and bootstrap procedures allow not only for compiling or averaging of the results obtained across the various repetitions of the analysis but also for assessment of the statistical significance of the results obtained at each resampling of the analysis. It is not unusual for results that were originally statistically significant to be statistically nonsignificant across at least some of the jackknife or bootstrap repetitions. Hence, Onwuegbuzie and Daniel coined the term "*Type V error*" to describe internal replication error rates, which provides information about how stable the computed *p*-value is across multiple resamplings of the same dataset. The internal replication Type I error rate can be determined by simply computing the percentage of resamplings in which a result that was statistically significant using the full sample becomes statistically nonsignificant. Similarly, the internal replication Type II error rate can be determined by computing the percentage of resamplings in which an originally statistically nonsignificant result with an acceptable level of statistical power yields a result outside the range of acceptable statistical power.

Type VI Error (Reliability Generalization Error)

It is important that researchers remember that the quality of their research studies is no better than the quality of the data they collect and analyze. Hence, it is always important to link statistical results to the measurement characteristics of scores (i.e., validity and reliability) on the measures used to generate those results. Questions about the validity of scores on one or more measures of variables used in a study can result in failure to effectively address a study=s research questions. Insufficient reliability evidence for data used to measure the variables results in increased error variance which diminishes the power of statistical results (Hays, 1988; Onwuegbuzie & Daniel, 2000a; Vockell & Asher, 1995). Further, low reliability coefficients for scores on one or more variables included in a study can create artificial "ceilings" for correlations between variables considering that the correlation correlation between any two variables cannot exceed the square root of the product of the reliabilities computed for the data on each variable (Crocker & Algina, 1986).

Reliability generalization is especially problematic considering that (a) relatively few researchers report reliability coefficients for the data they have collected within a given study (Onwuegbuzie, 1999; Vacha-Haase, 1998; Wilkinson & APA Task Force on Statistical Inference, 1999), (b) variations in sample composition can lead to variations in reliability coefficients for the same instrument (Vacha-Haase, Kogan, & Thompson, in press), and (c) homogeneity of scores can attenuate reliability estimates (Roberts & Onwuegbuzie, 2000), and (d) dependent variable scores in group comparison studies are subject to differential reliability estimates for subsamples (i.e., groups) within the data set (Onwuegbuzie & Daniel, 2000a). Onwuegbuzie and Daniel (2000a) proposed the use of an "upper bound" confidence interval when interpreting reliability coefficients. This confidence interval can be used to assess the degree to which the reliability coefficient derived with a given sample differs from an inducted-sample reliability coefficient (i.e., the reliability coefficient reported by the instrument's developers or in another research study).

Type VII Error (Heterogeneity of Variance/Heterogeneity of Regression Error)

Parametric statistical procedures rely upon certain assumptions about the nature of the data being analyzed which help assure that the procedures will appropriately honor the reality of the data. Among these assumptions are several requirements regarding homogeneity (e.g., the homogeneity of variance assumption in analysis of variance [ANOVA], the homogeneity of regression assumption in analysis of covariance [ANCOVA], and the homogeneity of treatment-difference variances assumption in multivariate analysis of variance [MANOVA]). The ANOVA homogeneity of variance assumption provides that the dependent variable variance will be approximately equal across the populations represented by the levels of the independent variable. There is some disagreement as to whether violations of this assumption are sufficient enough to cast doubt upon the legitimacy of the ANOVA results, with many holding the position that ANOVA is a relatively robust procedure when this assumption is not met prior to interpretation of group comparison statistics, particularly if the *n*' s of the groups being compared are equivalent (Heiman, 1996; Shavelson, 1996), though adjusted homogeneity *F* tests, such as the Welch, James, and Brown and Forsythe tests, are recommended by some (e.g., Elmore & Woehlke, 1996; Hinkle, Wiersma, & Jurs, 1998; Maxwell & Delaney, 1990) when the *n*' s are not equivalent.

A similar statistical assumption, the homogeneity of regression assumption in ANCOVA, tests the relationship between the covariate and the dependent variable across each level of the independent variable. This assumption is generally perceived to be extremely important because ANCOVA is <u>not</u> robust to this assumption' s violation (Hinkle et al., 1998). Failure to meet this assumption indicates that the covariate has unequally "adjusted" the dependent variable scores across groups. It is generally recommended not to continue with the ANCOVA analysis if the homogeneity of regression assumption is not met.

Various tests of the homogeneity assumptions have been proposed, with Levene' s (1960) *F* test among the most commonly employed procedures. The Levene test compares dependent variable variances or regression slopes across levels of the independent variable. A desirable Levene test result is a statistically nonsignificant *F* value, signifying that the variances or regression slopes are roughly equivalent and that the homogeneity assumption has been met. Checking for and tracking of Type VII error can be undertaken in several ways. First, the researcher should simply make sure that the appropriate homogeneity tests are conducted prior to employing ANOVA, ANCOVA, and other "OVA" procedures. Second, if multiple statistical tests are run, the researcher should keep track of the number of tests for which data either meet or fail to meet the assumptions (ANCOVA should be discontinued if the homogeneity of regression assumption is not met). Third, the researcher should be aware that homogeneity tests, like all statistical significance tests are subject to probability pyramiding (Stevens, 1996). Hence, the researcher should limit the number of ANOVA or ANCOVA tests (and concomitantly the number of homogeneity tests) run with a single data set and/or account for increased Type I error via use of conservative alpha levels and/or corrections to alpha.

An important assumption when conducting MANOVA procedures pertains to the homogeneity of treatment-difference variances (i.e., *sphericity*) assumption. This assumption requires that every measure must have the same variance and that all correlations between any pair of measures must be the same (Maxwell & Delaney, 1990). Because this assumption is very difficult to meet (Onwuegbuzie & Daniel, 2000b), in order to avoid committing a Type VII error, we recommend that researchers use the multivariate approach to analyzing repeated-measures data (which bases its analysis on difference scores) rather than the mixed-methods approach (i.e., with one factor representing the between-subjects factor(s) and the other factor representing the within-subject factor(s)) because the latter necessitates the sphericity assumption that is not required by the former.

Type VIII Error (Test Statistic Distribution Error)

This type of error, closely akin to Type III error, is introduced into analyses when the researcher expresses alternative hypotheses as directional yet assess results against a distribution of a test statistic as if he/she were using difference hypotheses. In terms of distribution of the test statistics associated with these hypotheses, this is essentially confusion of one-tailed and two-tailed tests. Recall that Type III error is related to the posing of an alternative hypothesis that predicts the directionality of the test in the incorrect direction. Type VIII error results when the researcher poses an alternative hypothesis in one direction or the other, but then utilizes the non-directional (two-tailed) portions of the distribution of the test statistic to test for statistical significance. Obviously, this type of error could also work in reverse (i.e., the researcher could pose a two-tailed hypothesis and then test only on one side of test statistic's distribution); however, this latter expression of the error is not likely considering that Type VIII error is most typically the result of statistics packages' use of the two-tailed test as a default.

Because parametric tests typically employ some sort of test statistic (e.g., *t*, *F*) for which a distribution is known, tests for statistical significance generally yield favorable (i.e., statistically significant) results when the obtained result is associated with test statistics toward the extremes (i.e., the tails) of the distributions. For example, for relatively large sample sizes (over 150), $t_{\text{CALCULATED}}$ values greater than a critical value of |1.96| are generally associated with statistical significance at the .05 level. This critical value, however is appropriate when one is testing a difference (two-tailed) hypothesis as the alternative. If the alternative hypothesis is stated in directional (one-tailed) terms, however, and alpha is kept at .05, it is appropriate to consider a difference only if it occurs at the particular tail of the distribution, and the critical value of *t* is reduced to either +1.645 or -1.645. Hence, it is easier to obtain statistical significance using a one-tailed (directional) test than it is using a two-tailed (difference, non-directional) test, assuming that the researcher is accurate in predicting directionality.

The problem with employing the incorrect portion(s) of the test statistic's distribution for testing the directional hypothesis, is that the researcher may fail to reject a null hypothesis that really is "rejectable" if the correct portion(s) of the distribution had been consulted. This would constitute what we are calling the Type VIII error. Type VIII errors are probably more common than might be immediately apparent due to researchers' tendency in many cases not to give the exact hypotheses that were tested (Hays, 1988; Johnson & Christensen, 2000).

Type IX Error (Sampling Bias Error)

By "Type IX error," we refer to as sampling bias error that results in inconsistency of results across studies. This is to some degree an extension of Type I error; however, it is more particularized in that it refers to disparities in results generated from numerous convenience samples across a multiplicity of similar studies. This type of error results in what is sometime erroneously referred to by descriptors such as "contradictory and inconclusive findings." These results are typically not contradictory or inconclusive, but rather represent differences in the populations being tested. Because convenience samples are used commonly in educational research, Type IX error is quite common.

In discussing result generalizability, Cook and Campbell (1979) and Pedhazur and Schmelkin (1991) distinguished between samples that allow one to generalize *to* a population of interest and those that allow one to generalize *across* populations of interest. Samples selected at random or by other systematic means from a population of interest allow for the former type of generalization. By contrast, because it is frequently less clear what populations they represent, convenience samples allow for the latter type of generalization. Type IX error is only error to the extent that researchers confuse the two types of research generalizations. Control of this type of error results from researchers' willingness to carefully consider the nature of the samples employed when interpreting research findings. Additionally, providing complete information about samples employed in research reports may serve to further quell this source of error.

Type X Error (Degrees of Freedom Error)

Type X error focuses on the tendency of researchers using certain statistical procedures (chiefly stepwise procedures) to erroneously compute the degrees of freedom utilized in these procedures. This final type of error in our proposed expanded typology is perhaps less visible than are other error types though equally deleterious in its effects on research quality. Degrees of freedom are essential to an understanding of most statistical procedures, representing the number of values within a set of observations that are free to vary under the assumption that the data in hand will yield the value of a given statistic in the population. In many statistical procedures, researchers expend degrees of freedom as they continue to make additional comparisons or compute correlations among additional variables. Rules for expending of degrees of freedom tend to follow logical patterns. For example, computation of a descriptive statistic, such as a mean, expends one degree of freedom; if two means are compared in a *t*-test, the researcher expends two degrees of freedom; if four variables are entered into a regression analysis, the researcher expends four degrees of freedom.

One glaring exception to this logical pattern is the calculation of degrees of freedom in stepwise analyses. While stepwise methods are problematic for many reasons (cf. Cliff, 1987; Huberty, 1994; Onwuegbuzie & Daniel, 2000b; Thompson, 1995), miscalculation of the degrees of freedom when determining variable entry order in procedures such as multiple linear regression analysis and discriminant analysis is particularly problematic. For a set of *k* predictor variables used in an analysis, stepwise procedures seek to determine which predictor variable serves as the best predictor of a given criterion variable. Determination is then made as to the predictor making the second best contribution to explaining the criterion once the variance explained by the first predictor is removed from the analysis, the third best contributor once the variance explained by the first two predictors is removed, and so on. An alternative to this "forward" stepwise procedure is the "backward elimination" method. Backward elimination puts all the predictors into the model and then removes them one at a time based on the variable that makes the least unique contribution to the analysis at any step of the procedure. Frequently, researchers utilize stepwise routines that combine both of these procedures (largely because this is the default for a stepwise routine in many statistical programs), resulting in a stepwise analysis that may alternately include or exclude variables at any step of the analysis based on the degree to which a previously-excluded variable makes a uniquely "new" contribution, or to which a previously-included variable is found to offer a primarily redundant contribution at the particular step of the analysis.

In computing degrees of freedom for testing stepwise hypotheses, researchers typically account for one degree of freedom for every variable actually entered into the predictive equation. However, it is important to note that variables that are consulted but not included at any step of the analysis should also be credited as expending one degree of freedom. The stepwise analysis has considered their input, deemed their contribution to be negligible or inappropriate at the given step of the analysis, and then assigned them a weight of zero. For example, in a four-predictor multiple linear regression case, at step one, the relationship of each predictor with the criterion is assessed even though only one variable is eventually included at that step, and even though most computer programs would only account for expending of one degree of freedom. In actuality, this step of the analysis should have expended four degrees of freedom. Hence, when one correctly computes the degrees of freedom for many stepwise analyses, the likelihood of obtaining a statistically significant result will diminish greatly.

An additional problem identified by the Onwuegbuzie and Daniel (2000b) is that because stepwise regression analyses utilize a series of statistical significance tests, these analyses are subject to actual Type I error rates that can be much greater than the nominal value of alpha. As noted by these authors, a stepwise regression procedure that takes five steps to select a final model, with the entry criterion being set at .05 (which is the default value for statistical packages), results in the probability of at least one Type I error rate being .23 (i.e., $1 - (1 - .05)^5$). Moreover, if some variables that are entered are then subsequently removed, then the Type I error rate can increase even more. Simply put, Type I errors that are inherent in stepwise procedures exacerbate Type X errors.

In quelling Type X error, we advocate that researchers seriously consider abandoning stepwise procedures altogether. This will not only eliminate the degrees of freedom problem mentioned herein, but will also serve to eliminate other deleterious problems related to these methods (Thompson, 1995). Further, when interpreting stepwise results presented by other researchers, it is important to determine what specific stepwise procedures were employed and to recompute degrees of freedom, if necessary, prior to accepting the researchers′ conclusions about statistical significance of the findings.

Conclusion

Without a doubt, social science research has been and continues to be plagued by a multitude of errors, misinterpretations, flaws, and shortcomings. Reflective researchers have historically called for better research practices via attempts to codify the types of problems they have found in research. The enduring system of identifying specific research errors and identifying them via Roman numerals has served effectively to allow researchers to have a common point of reference in discussing particulars errors that occur frequently and have appreciable effects on the quality of research interpretations. It is our hope that the expansion of this typology as we propose herein will serve to further the improvement of research methodology and practice.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Benton, R. L. (1991). Statistical power considerations in ANOVA. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (pp. 119B130). Greenwich, CT: JAI Press.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-536.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37,* 325-335.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 70*, 426-443.

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

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

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis for field settings*. Boston: Houghton Mifflin.

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

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. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2), 23-32.

Daniel, L. G. (1999, April). Assessing the quality of educational research: Issues and trends across a century of scholarship. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Daniel, L. G. (2000, January). Generalizability, replicability, and external validity in educational research: Sorting through terminology, exploring practical issues, and addressing common misconceptions. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Dodds, J. (1998). *Understanding interaction effects and Type IV errors*. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 426 094)

Elmore, P. B., & Woehlke, P. L. (1997). *Basic statistics*. New York: Longman.

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

Games, P. A. (1973). Type IV errors revisited. *Psychological Bulletin, 80*, 304-307.

Games, P. A. (1978). Nesting, crossing, Type IV errors, and the role of statistical models. *American Educational Research Journal, 15*, 253-258.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1988). *The empire of chance*. Cambridge: Cambridge University Press.

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

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed-effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-238.

Gold, D. (1969). Statistical tests and substantive significance. *American Sociologist, 4*, 42-46.

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.

Harwell, M. (1998). Misinterpreting interaction effects in analysis of variance. *Measurement and Evaluation in Counseling and Development, 31*, 125-135.

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Holt, Rinehart & Winston.

Heiman, G. W. (1996). *Basic statistics for the behavioral sciences* (2nd ed.). Boston: Houghton Mifflin.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston: Houghton Mifflin.

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.

Huck, S.W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.

Johnson, B., & Christensen, L. (2000). *Educational research: Quantitative and qualitative approaches*. Boston: Allyn and Bacon.

Kaiser, H. F. (1966). Directional statistical hypotheses. *Psychological Review, 67*, 160-167.

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.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donohue, 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.

Levene, H. (1960). Robust tests for equality of variance. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*. Palo Alto, CA: Stanford University Press.

Levin, J. R., & Marascuilo, L. A. (1972). Type IV errors and interactions. *Psychological Bulletin, 78*, 368-374.

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. *American Educational Research Journal, 7*, 397-421.

Marascuilo, L. A., & Levin, J. R. (1976). The simultaneous investigation of interaction and nested hypotheses in two-factor analysis of variance designs. *American Educational Research Journal, 13*, 61-65.

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.

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

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.

Onwuegbuzie, A. J. (1999, September). *Common analytical and interpretational errors in educational research*. Paper presented at the annual meeting of the European Conference on Educational Research, Lahti, Finland.

Onwuegbuzie, A. J. (2000a, November). *Effect sizes in qualitative research*. Paper to be presented at the annual meeting of the Association for the Advancement of Educational Research (AAER), Ponte Vedra, Florida.

Onwuegbuzie, A. J. (2000b, November). *Expanding the Framework of internal and external validity in quantitative research*. Paper to be presented at the annual meeting of the Association for the Advancement of Educational Research (AAER), Ponte Vedra, Florida.

Onwuegbuzie, A. J., & Daniel, L. G. (2000a, November). *Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Onwuegbuzie, A. J., & Daniel, L. G. (2000b, April). *Common analytical and interpretational errors in educational research*. Paper presented at the annual conference of the American Educational Research Association (AERA), New Orleans.

Onwuegbuzie, A.J., & Daniel, L.G. (in press). Uses and misuses of the correlation coefficient. *Research in the Schools*.

Oxford, R., & Daniel, L. G. (in press). Basic cross-validation: Using the holdout method to assess the generalizability of results. *Research in the Schools*.

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

Roberts, J. K., & Onwuegbuzie, A. J. (2000, November). *Alternative approaches for interpreting alpha with homogeneous subsamples*. Paper presented at the annual conference of the Mid-South Educational Research Association, Bowling Green, KY.

Rozeboom, W. M. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Boston: Allyn and Bacon.

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

Thompson, B. (1989). Asking "what if" questions about significance tests. *Measurement and Evaluation in Counseling and Development, 22*, 66-67.

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

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

Umesh, U. N., Peterson, R. A., McCann-Nelson, M., & Vaidyanathan, R. (1999). Type IV error in marketing research: The investigation of ANOVA interactions. *Journal of the Academy of Marketing Science, 24*, 17-26.

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. (in press). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*.

Vockell, E. L., & Asher, J. W. (1995). *Educational research* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

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.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM032244

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Towards an Extended Typology of Research Errors

Author(s): Larry G. Daniel and Anthony J. Onwuegbuzie

Corporate Source: Univ. of North Florida

Publication Date: November 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Larry G. Daniel Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Larry G. Daniel

Printed Name/Position/Title: Larry G. Daniel, Assoc. Dean

Organization/Address:
Univ. of North Florida, COEHS
Jacksonville, FL 32224-2676

Telephone: (904) 620-2520

FAX: (904) 620-2522

E-Mail Address: ldaniel@unf.edu

Date: 11/17/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com