QUEST, 2001, **53**, 418-439

Revisiting the Assumptions for Inferential Statistical Analyses: A Conceptual Guide

Ang Chen and Weimo Zhu

Trustworthiness of results from statistical analyses relies on the fulfillment of a set of assumptions made about data. A survey of published pedagogy studies in physical education revealed that examining the assumptions has been overlooked in data analyses and result reporting. The purpose of this article is to provide a conceptual understanding of the assumptions and to summarize available methods to test and address their violations. In the article, we present information to show that it is inappropriate to overlook the assumptions and ignore their violation's impact on interpretation of results. We summarize current findings from theoretical statistical research that start to challenge the conventional belief about the robustness of traditional inferential statistical analyses. Remedial procedures recommended in the latest statistical theories are presented for researchers to adopt in order to generate valid results from statistical analyses.

Trustworthiness of results from statistical analyses relies on the fulfillment of a set of assumptions made about data condition. When the assumptions remain unknown in research reports, it is very difficult for research consumers to judge the validity of the results. As reported recently (Keselman et al., 1998), many educational researchers have overlooked examination of the assumptions or have not viewed it as an important part of data analysis and reporting. At times, published data clearly showed violation of the assumptions and/or unawareness of researchers about the violations. Conducting statistical analyses without examining data condition is problematic (Huck, 2000; Lomax, 1998). At best, it results in findings that are not replicable and generalizable. At worst, it generates misleading educational policies and curricular/instructional decisions that jeopardize the future of young generations.

This article focuses on issues of assumption examination involved in research that adopts statistical analyses and uses statistics as the primary tool in making inferential decisions about the data. Our purpose is to assist readers to

Ang Chen is with the Department of Kinesiology at the University of Maryland at College Park, MD 20742. E-mail: <ac192@umail.umd.edu>. Weimo Zhu is with the Department of Kinesiology at the University of Illinois at Urbana-Champaign.

conceptually understand the assumptions, options to test them, and strategies to address their violations. We do not intend to focus on computational techniques. For these techniques, the reader is encouraged to consult the references we cite and similar articles with excellent examples from our fields (e.g., Thomas, Nelson, & Thomas, 1999). After a brief overview of current indicators of whether and how assumptions are examined in physical education research, we will focus on discussing types of assumptions, the importance and consequences of violations, approaches to assumption examinations, and alternative statistical analysis methods when the assumptions are violated.

A Summary of Pedagogy Research in the 1990s

To understand the current practice in using inferential statistical analyses in physical education research, we surveyed pedagogy research articles published from 1990 to 1999 in *Research Quarterly for Exercise and Sport* and *Journal of Teaching in Physical Education*. We believe that these journals are accessible for most pedagogy researchers in physical education in North America. For article selection, we defined pedagogy research as empirical inquiries that address issues directly related to teaching and learning in school settings (elementary and secondary schools and colleges) and involve teaching/learning variables.

We identified 295 research articles, including 122 (41%) qualitative inquiries and 173 (59%) quantitative studies. As can be seen in Table 1, 125 of the quantitative studies used and 48 did not use inferential statistical analysis. It seems that quantitative method is a major tool in pedagogy research, and researchers rely on inferential statistics to reach conclusions. Our overview was based on the 125 articles that involved inferential statistical analysis.

The 125 studies represent a wide array of research designs and statistical analysis strategies. The sample sizes ranged from 9 to 1,371. The mean sample size is 175 ($SD = 247$) and the median is 74. Approximately 20% of the studies used samples of less than 25 participants. The mean number of dependent variables is 10 ($SD = 14$) and the median is 5. About 30% of studies involved more than 10 dependent variables. For many studies, the main purpose was to develop an instrument to measure a phenomenon in the teaching/learning process in physical education. In these studies, responses were analyzed usually on the item basis. In most of these situations, multiple mean comparisons (e.g., multiple t-test, multiple ANOVA

Table 1 Summary of Research Type by Journals

Journal	Total	Qualitative	Inferential Statistics	Descriptive Statistics
Research Quarterly for Exercise and Sport	44	19 (43%)	21 (48%)	4 (9%)
Journal of Teaching in Physical Education	251	103 (41%)	104 (41%)	44 (18%)
Total	295	122 (41%)	125 (43%)	48 (16%)

one-way analysis) were conducted without adequate adjustment to control possible inflation in actual α. Most studies (70%) involved one independent variable, another 20% used two. The remaining 10% of studies examined more than three independent variables and one study involved eight independent variables. Most studies (77%) compared differences among the means using inferential statistical analyses (*t*-test, ANOVA, ANCOVA, MANOVA, etc.). Correlational data analyses (correlation, factor analyses and regression) were used in about 13% of the studies. About 10% used non-parametric data analysis strategies, such as χ^2, U-test.

Assumption examinations were reported in 12 (9%) of the 125 studies. This is consistent with the claim that the examination on data conditions has not been a major concern in data analysis for educational researchers in the past 10 years (Keselman et al., 1998). Table 2 lists these 12 studies and summarizes the types of statistical analyses used, types of assumptions examined, and whether results of the examinations were reported.

A striking fact revealed in Table 2 is that assumptions were examined in only a fraction (9%) of the studies that used inferential statistical analyses. Although this fact cannot be taken as suggesting a high rate of assumption violation, it certainly presents an ambiguous picture to research consumers in physical education. It can be assumed that there are no less assumption violations in these studies than those found in the educational research literature (Keselman et al., 1998; Lix, Keselman, & Keselman, 1996). For instance, we found that several studies used Likert-Type scale (1-5) in measuring dependent variables and reported means of greater than 4.0 and standard deviations of greater than 1.0. In these cases, the normality assumption of the data becomes questionable. When inferential statistical analyses conducted without reconditioning the data (as were in these studies), the results could be misleading (Wilcox, 1998).

The Primary Assumptions

There are three primary assumptions related to inferential statistics: observation independence, normality of frequency distribution, and equal variance. Some concerns for violations of the assumptions can be addressed in the research design phase. Others should be examined after data are collected but before inferential statistical analyses. In the following discussion, we focus on the latter but will address the former when necessary. In addition, our discussion will be mainly in the realm of between- and within-subject designs to which *t*-test, ANOVA, and MANOVA are often applied as major data analysis tools. But the reader should be aware that these assumptions also govern correlational analyses.

Observation Independence

The observation independence assumption states that dependent measures must be taken independently among the participants in the sample. The measures must be independent both within and between treatment/control or different observation groups. Independence means that regardless of measurement methods, each measurement/observation on an individual participant is in no way related to the same measurement/observation made on another participant. In statistical terms, the independence assumption is defined as independence of residuals, expressed as error terms in inferential statistics (i.e., ϵ [in ANOVA, MANOVA]). The

Table 2 Summary of Assumption Tests

Statistical Analysis	Assumption Examined	Result Reported
Journal of Teaching in Physical Education (n = 104)		
1. ANCOVA	Homogeneity of covariance	Box M's F value reported, no df and p
2. MANOVA	Homogeneity of covariance, Sphericity	Values not reported
3. ANOVA	Homogeneity/Linearity of Variance	Results reported
4. ANOVA	Homogeneity of variance, Sphericity	Values not reported
5. ANOVA	Normality	Violation of assumptions was described in text and data were transformed. But types of tests and test results (e.g., skewness) were not reported.
6. Q analysis	Normality is assumed in the design of data collection method.	Described in text.
7. MANOVA	Sphericity	Not reported
8. Repeated MANOVA	Multicollinearity	No explanation provided as to why this assumption was examined.
Research Quarterly for Exercise and Sport (n = 21)		
9. MANOVA	Homogeneity of covariance, Sphericity	Values reported, violation explained, adjustment made
10. Factor Analysis	Normality	Reported
11. ANCOVA	Homogeneity of covariance	Not reported
12. Multiple ANOVA	Independence	Violations reported, adjustments on means made

Note. n = total number of reports published in the journal that involved statistical analyses.

assumption of independence is satisfied when the distribution of ϵ on each measure is randomly distributed as illustrated in Figure 1(a). One can suspect a violation of the assumption when the distribution is in a nonrandom pattern as shown in Figure 1(b).

Consequence of Violation. Observation independence is the most important assumption (Huck, 2000; Keppel, 1991; Lomax, 1998; Stevens 1990, 1992). Its violation will mislead interpretations of research results. As Keppel (1991) argues, when measurements are not taken independently, there is always a confounding of variables. Consequently, the researcher is unable to logically make inferences about

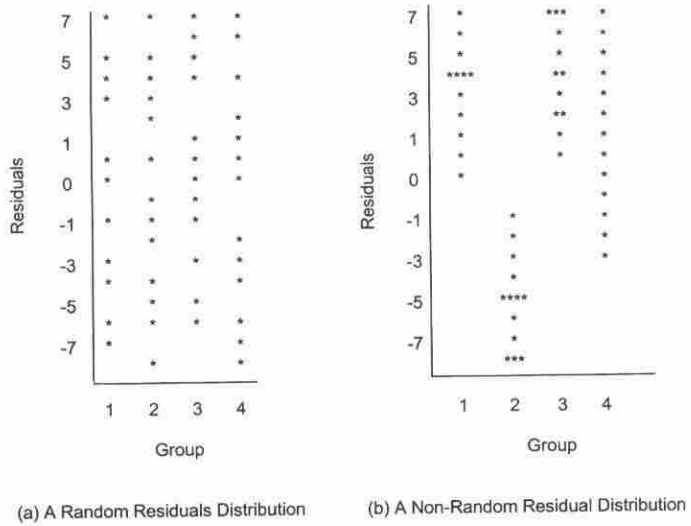


Figure 1 — Random and nonrandom residual plots.

the effects from the independent variables on the dependent variables. Statistically, the violation produces a substantial effect on the level of statistical significance. Scariano and Davenport (1987) demonstrated when the assumption is violated a p value of .05 is actually .54 (3 groups of 10 participants each with a moderate intracorrelation of .30), 10 times greater than the p value appears to be! Scariano and Davenport also showed that increasing sample size does not help. In fact, when the assumption of independence is violated, the larger the sample is, the more severe the consequence will be. Figure 2 shows the joint influence of autocorrelation (intra-correlation, ρ) and sample size (N) on Type I error rate (Huitema, McKean, & McKnight, 1999). Adopted from Scariano and Davenport (1987), Table 3 demonstrates the relationship between correlated observations (intraclass correlation coefficients) and actual Type I error rates (actual p values).

Possible Causes. It is believed that the violation of observation independence occurs in three measurement contexts (Lomax, 1998). The first is the repeated measurement context. This context includes a pre- and post-measurement design, or a multiple-measurement design, or a longitudinal design. In these research designs, one or more dependent variables are measured multiple times over time with nondistinguishable treatments. In other words, time is used as the sole independent variable. When a variable is measured multiple times over a period, the two measures taken in closely adjacent points of time are more likely to have an auto correlation than any other pairs measured with a relatively larger time interval.

The second context is to measure dependent variables within an intact grouping setting (within blocks) such as in intact physical education classes. In this context, when the dependent measures are influenced by a unique factor or factors that is different from that in another group, it is likely that responses from each group are auto correlated within the group.

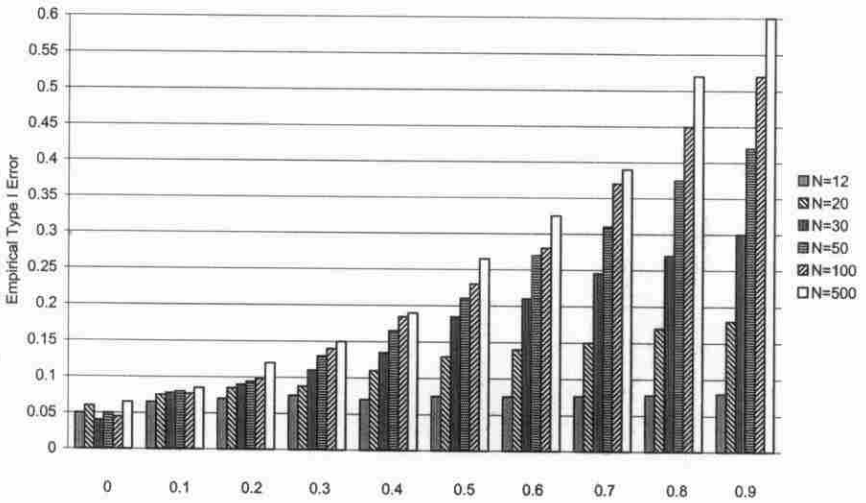


Figure 2 — Joint influence of ρ and N on Type I Error Rate (Adopted from Huitema et al., 1999).

Table 3 Actual Type I Error Probability (p) With Change of Autocorrelation (ρ) When α Is Set at .05 (Scariano & Davenport, 1987)

Group	n	ρ = .00	ρ = .01	ρ = .10	ρ = .30	ρ = .50	ρ = .70	ρ = .90	ρ = .95	ρ = .99
2	3	.05	.05	.07	.14	.24	.38	.63	.74	.88
	10	.05	.06	.17	.37	.53	.68	.83	.88	.95
	30	.05	.08	.34	.59	.72	.81	.90	.93	.97
	100	.05	.17	.57	.77	.75	.90	.95	.96	.98
3	3	.05	.05	.08	.17	.34	.56	.84	.92	.98
	10	.05	.06	.22	.54	.73	.87	.96	.98	.99
	30	.05	.10	.49	.80	.91	.96	.99	.99	.99
	100	.05	.22	.78	.93	.97	.99	.99	.99	.99
5	3	.05	.05	.10	.27	.52	.78	.97	.99	.99
	10	.05	.07	.32	.75	.92	.98	.99	.99	1.00
	30	.05	.12	.69	.95	.99	.99	.99	1.00	1.00
	100	.05	.32	.94	.99	.99	.99	1.00	1.00	1.00
10	3	.05	.06	.13	.45	.78	.97	.99	1.00	1.00
	10	.05	.08	.50	.94	.99	.99	1.00	1.00	1.00
	30	.05	.16	.91	.99	1.00	1.00	1.00	1.00	1.00
	100	.05	.49	.99	1.00	1.00	1.00	1.00	1.00	1.00

Note. Actual *p* values are rounded up from 4-digit values on the original table.

The third context is to replicate a study with an identical sample within a short time frame. A hypothetical example of such cases could be that a researcher wants to study the effects of a new teaching method on learning a movement skill in students with a certain type of disability. In such a situation, only a limited number of students may be available for a series of replication experiments. The researcher replicates the study several times with the same group of students without extended time intervals. As a result, the measurements taken in the later experiments are not independent from the effects of the previous ones.

Testing the Assumption. In the case that a study is not intervention in nature and that the tenability of observation independence is not addressed in research design, it becomes necessary to assess auto correlation in the dependent measures. This assessment must be conducted prior to the inferential statistical analyses because it provides crucial information for the researcher to make knowledgeable decisions on whether or not to apply conventional inferential statistics in the data analysis. The common indicator of auto-correlation is intracorrelation coefficient computed using $\rho = (MS_b - MS_w)/(MS_b + (n - 1)MS_w)$, where

ρ: intra-correlation coefficient

MS_b: between-group mean square (numerator of the F statistic)

MS_w: within-group mean square (denominator of the F statistic)

n: number of observations in each group.

In most statistical software packages, the statistics can be obtained as intermediate calculation outcomes of F tests (e.g., ANOVA, ANCOVA, MANOVA).

Another approach to the assessment is to graph residuals of the dependent measures for each group of participants. As shown in Figure 1, the nonlinear, randomly distributed scores in a residual graph indicate a minimal risk of violating the assumption. Clustered scores, either positive (above 0) or negative (below 0), are indicative of a violation.

Approaches to Addressing the Assumption. To satisfy the assumption, it is important for researchers to address fundamental issues in research design. The issues include randomizing participants and/or treatments and planning data collection strategies for taking measurements individually. Pedagogy researchers often find it extremely difficult to have true random student samples and to collect data on an individual basis. In most school-based research, observations and measurements have to be conducted in intact classes in order for the researcher to comply with participating schools' class schedule. This reality presents an increased risk of violating the independence assumption. To address this issue in the research design phase, researchers can first use randomization in sampling. But in a multi-level system, which involves multiple level units such as school districts, schools, classes, and teachers and students, true random assignment is very difficult. Researchers should decide the scope of randomization in terms of the extent to which they intend to generalize the results of the investigation (Borg & Gall, 1989).

Given that true randomization is difficult to achieve, a viable choice is careful and strategic use of quasi-experimental design. The central issue of the independence assumption is to minimize auto correlation (represented by intracorrelation

statistic) of each dependent measure among participants within a group. When sampling, the researcher can look for research sites that involve minimal possibilities of long-term intragroup interactions on variables of interest. For example, the researcher may select participants in schools where the physical education program and class schedules are designed using a rotation system in which students do not necessarily take all the instructional units in intact classes taught by the same teacher.

The previous two approaches may help minimize the possibility of violating the assumption in the research design phase. Once the violation is suspected in the data analysis phase, the researcher is expected to focus on changing data conditions to satisfy the assumption before engaging in further statistical analyses.

One remedy is to recondition the data by redefining the analytical unit for statistical analyses. Silverman and Solmon (1998) have recommended that when within-group auto correlation is suspected, the researcher should use group means, rather than individual scores, as the unit of analysis. However, using group means will drastically reduce the sample size (N) and reduce the statistical power considerably, especially when other assumptions are violated also. Thus, using group means may increase the probability of Type II error, accepting a false H_0, and mask the difference in dependent measures that should be detected.

In nonintervention research, the decision on unit of analysis should be made on the basis of statistical testing results of intracorrelation and intercorrelation. Currently, the extent of negative effect of auto correlation is still in debate among theoretical statisticians. At least one recent report (Huitema et al., 1999) suggests that distortion of Type I error caused by auto correlation may be far less than that which has been predicted, especially when the intracorrelation (ρ) is less than .10 (see Figure 2). Therefore, detecting severity of auto correlation and assessing the consequence become essential tasks and should be addressed prior to determining the unit of analysis in inferential statistical analyses.

Another approach is to adjust the α level to base inferential decisions on the extent of auto correlation (Stevens, 1990, 1992). In this case, an intracorrelation coefficient (ρ) should be calculated and an appropriate p value should be selected in terms of the true Type I and Type II error probabilities. Scariano and Davenport (1987) have provided a set of calculated values for such adjustments. In a case where there are 5 groups of 30 observations each and ρ is .10, the Type I error rate will be approximately 14 times greater (p = .69 when it appears as .05). Thus, when the p value is .01, the researcher must be aware that the actual p value is .10 or greater, depending on the magnitude of the auto correlation. In such a case, a more stringent α level (.001 or smaller) should be used for a claim of p = .01. Figure 3 is a conceptual map summarizing the above approaches to addressing the assumption of independence.

A different solution to the problem is to adopt the hierarchical (multi-level) linear modeling (HLM) design in school-based research. HLM design allows researchers to analyze data taken from individual, group, and institution levels. Its data analysis methods prevent severe distortion of data caused by using conventional statistical analyses when the assumption is violated. Major HLM concepts and techniques have been introduced in our field (Zhu, 1997a). More detailed information can be found in Bryk and Raudenbush (1992), Heck and Thomas (2000), Lee (2000), Snijders and Bosker (1999), and Kreft and de Leeuw (1998).

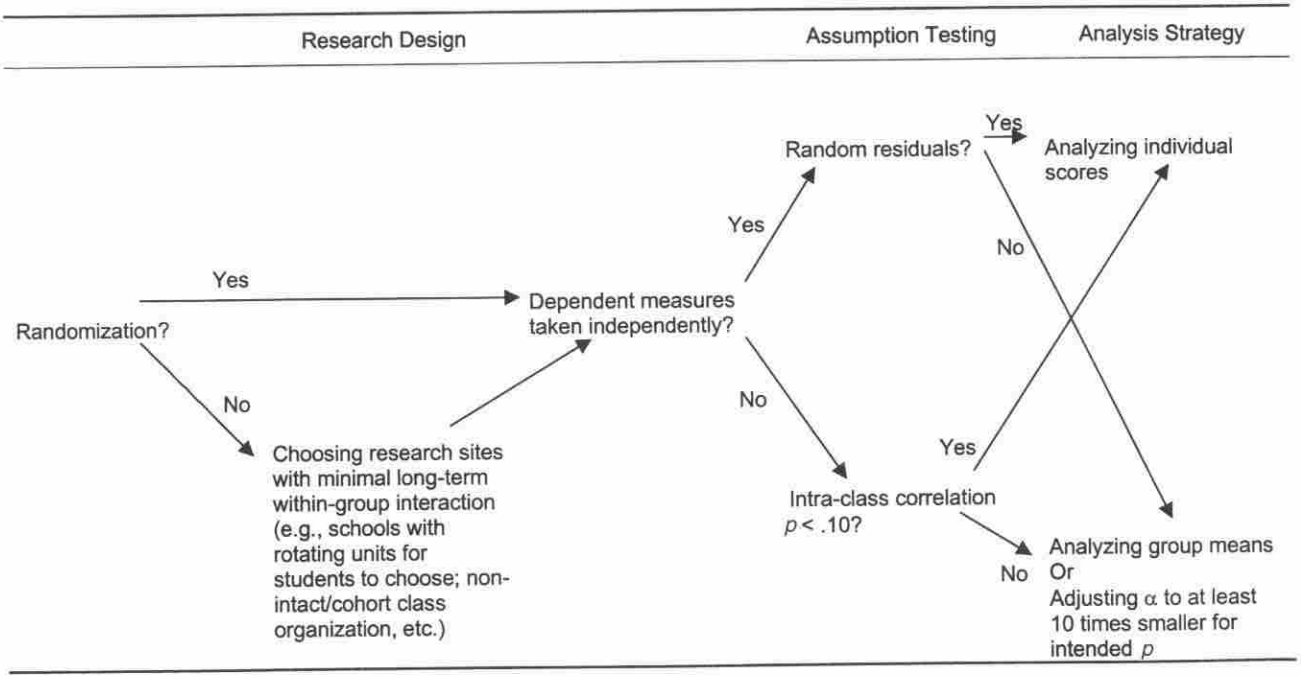


Figure 3 — A conceptual map for addressing assumption of independence.

Distribution Normality

Distribution normality of dependent measures states that the dependent measures are distributed in a normal bell shape that represents a standardized relationship between central tendency and variability of the measures. The concept of normal distribution differs from that of symmetry. A symmetrically distributed data set shares the bell-shaped graphical feature of a normal distribution but not its statistical characteristics.

Distribution normality is the very basis on which most inferential statistical tests, such as *t* and *F* tests, were developed (Micceri, 1989). In inferential statistical terms, the assumption requires a normal distribution of residual errors for the dependent measure within each cell (levels) of all independent variables (e.g., within male and female groups or elementary and middle school groups). However, with the belief of "robustness" in some approaches, such as *F* tests, educational researchers consider that the assumption is satisfied as long as the statistics examined in research reflect the parameters that have been assumed normally distributed in the target population (Geary, 1947, as cited in Micceri, 1989).

Consequence of Violation. When the assumption of normality is violated, the consequences can be in two directions. First, when the distribution is asymmetrical, an inflated Type I error rate may result. Another consequence is an escalated Type II error rate, accepting a false null hypothesis. Many analytical approaches, including analyses of variance, can have relatively low power when the data distribution departs from normality even slightly. Wilcox (1998) has demonstrated how power can decrease in nonnormally distributed data. In the context of Student *t*-test, a slight departure from normality with a group sample size of 25 can lower statistical power from .96 to .28 with an α of .05. Wilcox has speculated that many significant findings were lost because reduced power prevented many researchers from discovering differences with statistical significance.

Robustness Against Nonnormality. It is a widely held belief that *F* tests have a high level of robustness. In these analyses, violation of the normality assumption does not result in severe consequence in terms of Type I and Type II error rates. This belief is supported by a pioneer review on statistical research on robustness (Glass, Peckham, & Sanders, 1972) and has been acknowledged in most statistical textbooks (e.g., Lomax, 1998; Stevens, 1990).

The belief about robustness is based on the results of theoretical statistical (Monte Carlo) studies where simulated data are used to examine possible consequences of the violation. Clinch and Keselman (1982) and Tan (1982), however, have shown that conclusions about the robustness of *F* tests from Monte Carlo studies are conditional. Specifically, for *F* tests to provide robust results, the data should meet the following conditions: (a) the distribution of data has to be symmetrical within groups, (b) groups are equal in size, and (c) group size should be large. There is a disagreement about the group size. According to Clinch and Keselman (1982), observations in each group must be greater than 12. Others have recommended that a reasonable group size should be at least 20–50 (Bock, 1975) or even larger (Bradley, 1980). When these conditions are not met, robustness of *F* tests decreases.

On the other hand, as demonstrated by Micceri (1989), discrepancy between simulated data used in Monte Carlo studies and real-world data can be dramatic. Micceri found that data conditions characterized in most Monte Carlo studies were

rarely seen in the 440 data sets that he collected from various educational studies. The inconsistency has cast clouds over the validity of Monte Carlo study results and the belief of robustness.

Testing the Assumption. Several methods can be used to examine distribution normality of data. One is to use graphical techniques. Researchers can use stem-and-leaf plots or histograms to examine symmetry of raw scores and identify unusual scores (outliers). Another graphical technique is to draw a normal probability plot. For a normally distributed data set, the points on the plot will fall along a straight diagonal line. Researchers should be advised, though, that there is no criterion to help judge the departure of the points from the linearity. Researchers also can calculate the skewness statistic. When the skewness statistic is greater than 1.5, the data distribution can be considered as severely asymmetric. It is recommended to combine graphical and statistical techniques when testing the assumption (Lomax, 1998).

Addressing Violations of the Assumption. To prevent an inflated Type I error rate when the assumption is violated, Keppel (1991) suggests simply using a more stringent α criterion such as .025 or .01 in place of .05. However, when the violation is coupled with (a) small sample size ($n < 25$); (b) unbalance design (unequal n in groups, ratios between largest n to smallest $n > 1.5$); and (c) violation of the equal variance assumption, many theoretical statisticians (e.g., Lix et al., 1996; Lomax, 1998) generally agree that researchers should consider using alternative data analysis strategies to address violation of the normality assumption.

A common strategy is to transform data. But transformation must be used as the last resort because of the problems associated with selection of transformation indexes and their interpretation. For example, one of the most popular transformations is log transformation. As Pearson and Please (1975) have pointed out, when transforming the results back to an antilog mean of the measure after the analysis on the transformed log mean, the researcher should not interpret the antilog mean as the mean of the measure, because it is not.

In general, the conventional F tests are robust against a slight departure from normality. When the violation of normality assumption is not severe, and the equal variance assumption is met, the researcher can proceed with F tests. However, when the equal variance assumption is violated, the researcher should pursue alternative data analysis techniques (Lix et al., 1996; Lix & Keselman, 1998; Micceri, 1989; Wilcox, 1998).

One alternative strategy is to use nonparametric methods (Thomas et al., 1999). Nonparametric data analysis strategies do not rely on the assumptions of distribution normality and equal variance (discussed in the next section). The assumption for nonparametric analyses only requires independence of observations. In addition to the χ^2 method, one popular nonparametric alternative to ANOVA is Kruskal/Wallis procedure (1952). Thomas et al. (1999) have provided an excellent tutorial for using nonparametric statistics with computational examples from exercise science. These procedures are not difficult to understand and use. They are available in most statistical packages.

Although nonparametric methods can help avoid violation of the normality assumption, they are, in general, less powerful in detecting differences. Consequently, they demand large sample sizes in order to yield statistically significant differences. In addition, when using nonparametric methods, researchers must be aware that they are testing a hypothesis different from that of a parametric

method (such as those in *t*- and *F* tests). Therefore, interpretations of the results should be configured with the nonparametric methods, not with the original intention of the research (Lix et al., 1996).

To overcome this complication with nonparametric methods, several theoretical statisticians have proposed alternative techniques that allow researchers to adjust data conditions for parametric analyses. One such technique is trimming (Wilcox, 1998; Yuen, 1974). In trimming, a certain percentage (10% – 20%) of the largest and smallest scores is eliminated from the sample. The mean of the remaining scores (called trimmed mean) can be used. The statistic for variance associated with trimmed mean for variability is Winsorized variance. Yuen (1974) has provided a theoretical explanation of the two statistics and Wilcox (1996, 1997) has provided detailed explanations and statistical software for computing them. To compare the trimmed means, Lix and Keselman (1998) have recommended using Alexander and Govern (1994), James (1951), and Welch (1951) statistics rather than the conventional mean comparison tests.

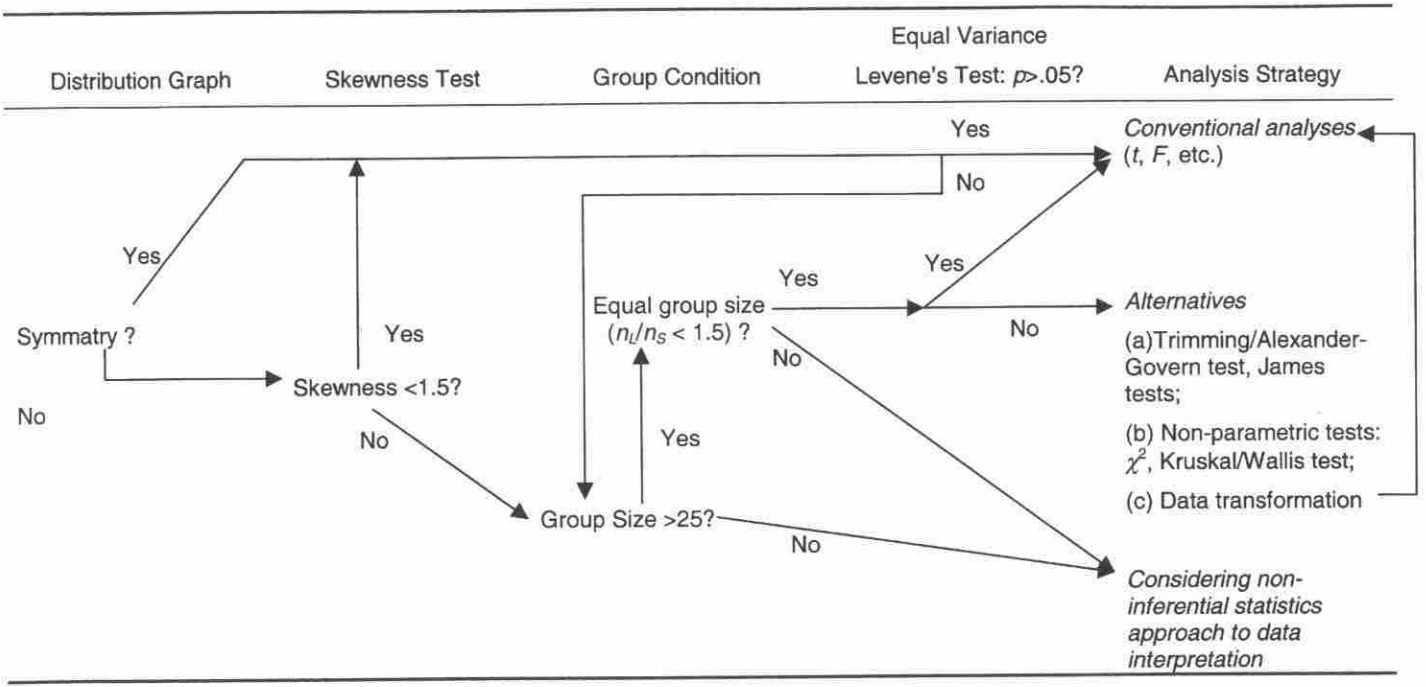
Finally, a set of computer-based resampling methods, known as the computer intensive methods, has become available for analyzing the data that violate the normality assumption. The methods artificially reconstruct the original data sets many times and compute and recompute the inferential statistics on these reconstructed data sets. This group of methods includes bootstrapping, Monte Carlo sampling, and approximate randomization test. Interested readers can refer to Noreen (1989) and Hjorth (1994) for a general description of these methods and Zhu (1997b) for application of the bootstrapping techniques in our fields. In Figure 4 we provide a conceptual map to describe the procedure for examining and testing violations and summarize the alternative strategies for data analysis when the assumption is violated.

Equal Variance

Equal variance (homogeneity of variance) is another important assumption in inferential statistical analyses. It states that the variance (s^2) of the dependent measure is equal among populations involved in a study. Generally, the assumption requires the variance (s^2) among all sample groups to be equal. Consequences of violation, however, may vary in terms of magnitudes of variances of the variables involved and the ratio of group sizes.

There are dozens of ways to statistically test the assumption of equal variance in one-way or factorial design. Among them include Hartley's F_{max} statistic (largest s^2/smallest s^2, when the ratio is greater than 3, the assumption is violated) when group sample sizes are equal, and the Bartlett test when group sample sizes are not equal (Kirk, 1982). All the tests are sensitive to the assumption of normality. In other words, when the data are not normally distributed, the power of detecting the difference in variance is affected.

Conover, Johnson, and Johnson (1981) examined 56 statistical tests for equality of variance and two were found to have worked best. One is the Brown-Forsythe procedure (Brown & Forsythe, 1974). This approach is based on testing transformed scores associated with the median, rather than the mean. Therefore, it is insensitive to data distribution. Another similar approach is Levene's test (Levene, 1960). These tests are incorporated in several statistical packages (e.g., SPSS) for users to specify.



Note: n_L = Largest group size; n_S = Smallest group size.

Figure 4 — A conceptual map for addressing assumption of normality.

Consequence of Violation. This assumption is considered more important than the normality assumption in that it directly affects the calculated F value (Keppel, 1991; Lomax, 1998; Stevens, 1990; Wilcox, 1997). A violation of this assumption affects the calculated value of within-group sum of squares (SS_{within}). SS_{within} is used to calculate the within-group mean sum of squares (MS_{within}); and MS_{within}, in turn, is used as the denominator in calculation of the F value. Therefore, the violation may lead to a biased F value and result in an increase in the Type I or Type II error rate (Lomax, 1998).

It has been believed that as long as group sizes are equal (largest/smallest < 1.5), F tests are relatively insensitive to the violation of the assumption. In other words, F tests are robust. When group sizes are not equal (largest/smallest >1.5), however, using the F tests will lead to a liberal F value causing an increase in Type I error rate (e.g., $p = .05$ may become $p = .48$). But the damage may be controlled by using a more stringent α as the critical criterion, such as $p = .01$ in place of $p = .05$, as long as the assumption of normality is met.

Recent developments in theoretical statistics have shown that the robustness of F tests is conditional. For example, Wilcox (1987) has shown that when the ratio of largest to smallest within-group variance (largest s^2/smallest s^2) is equal to or greater than 9, the computed F value can be seriously biased. Box (1954) has suggested that when the ratio of the largest and smallest within group variances is equal or greater than 3 (largest s^2/smallest $s^2 < 3$), the researcher should consider using alternative approaches for data analysis. In educational research, however, the variance ratios sometimes can be as large as or greater than 16 (Keppel, 1991), much larger than what can be tolerated. Relying on the conventional belief of robustness in F tests can be a problematic approach to data analysis.

Addressing Violations of the Assumption. When the equal variance assumption is violated, there are two often-used alternatives: (a) to transform data and proceed with F tests and (b) to adjust α levels to control inflated Type I error rates. However, as stated above, interpreting the meaning of transformed data is always difficult. In addition, because transformations (logarithm, square-root, etc.) should be based on the particular type and degree of the violation, it is not always an easy task for researchers to select an appropriate data transformation (Oshima & Algina, 1992). Adjusting the α level seems to be a weak solution because it is at the expense of statistical power. When the violation of equal variance is complicated with the violation of the normality assumption and unequal sample sizes (largest/smallest group > 1.5), the validity of these approaches becomes questionable (Wilcox, 1987). In that case, researchers are advised to use alternative statistical methods to control the Type I error rate as well as to maintain statistical power.

Several alternative methods can be used in such a situation. As summarized by Keppel (1991), the researcher could consider the Welch W test (Welch, 1951), Brown-Forsythe F^* test (Brown & Forsythe, 1974), James' second-order method (James, 1951), and a two-stage method proposed by Bishop and Dudewicz (1978).

Lix et al. (1996) reviewed alternative statistical methods including those mentioned above and the Kruskal-Wallis procedure, a nonparametric method, and they took into account violations of both normality and equal variance assumptions and compared the alternatives with F tests. The data reviewed were from 37 field-based educational studies rather than simulations. The criterion of robustness was set $.025 < \alpha < .075$, when the α was expected to be equal to .05. An important conclusion from the review was that all alternative methods performed

better than the ANOVA *F* test. Lix et al. (1996), therefore, suggest that when the assumptions are violated, the researcher should not use ANOVA *F* test. Among the parametric alternative methods, James (1951) and Welch (1951) tests performed best. However, they noticed that the robustness of the methods depends on (a) that nonnormality is not substantial (skewness < 2.0) and (b) that total group size in both balanced and unbalanced designs should be greater than 10. The nonparametric method, Kruskal-Wallis procedure, was found sensitive to the violation of equal variance assumption although it is robust against nonnormality. Lix et al. (1996) advise researchers to take extreme caution when using this method.

As Micceri (1989) noticed, the data from educational studies usually cannot meet the minimal standard for using the alternative methods in many instances. For example, the group sizes are not equal, the variances among the groups are not equal, or the data are skewed (skewness value > 2.0). "Under such conditions, researchers should consider other techniques for communicating research findings in order to avoid filling the literature with false positive results from invalid inferential procedures" (Lix, et al., 1996, p. 614).

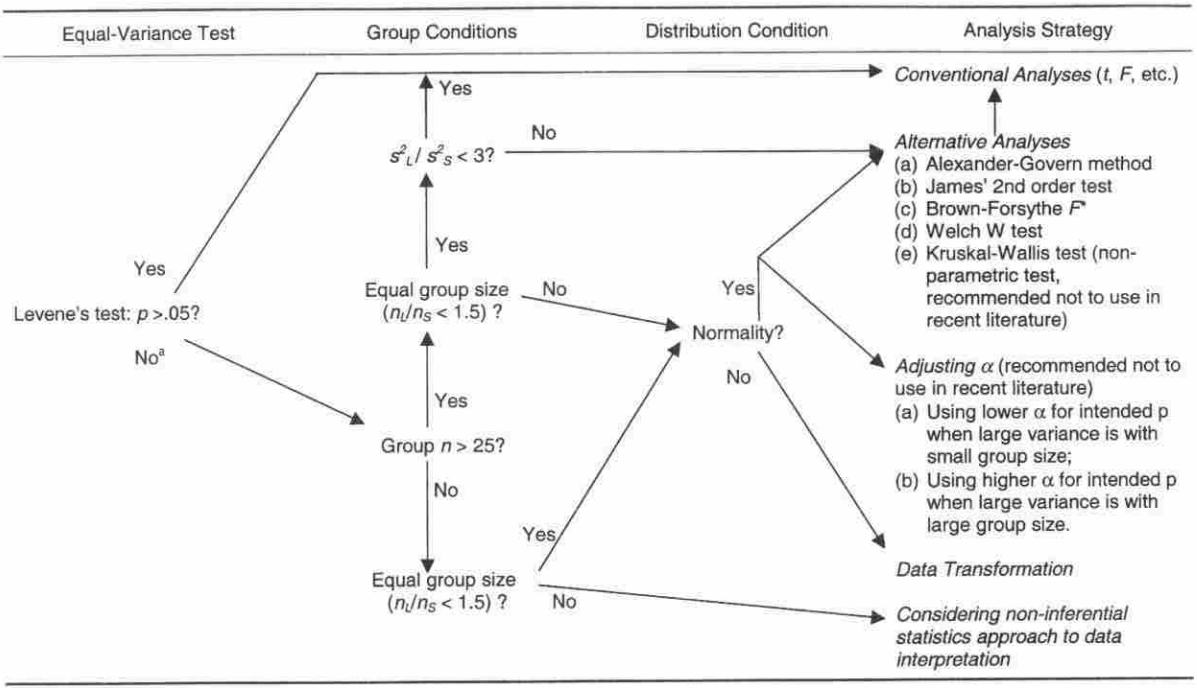
The available techniques may include trimming (as mentioned above), interpreting results based on central tendency and variability of data, and designing multistage, multidata studies to replicate the investigation. As Keppel (1991) has argued, statistical analysis is merely a tool to help researchers decide whether the results from a single study are dependable, in the sense that the results can be observed repeatedly. Replication is the core meaning of generalization. "If we find the same results on two independent occasions, we in essence validate the initial outcome of the study, which was somewhat in doubt because of the excessive Type I error" (Keppel, 1991, p. 108). As a summary, the conceptual map in Figure 5 describes the procedure for examining the assumptions and possible alternative data analysis methods available for addressing the violation of the assumption.

Additional Assumptions for Special Factorial Designs

ANCOVA. There are instances when the researcher is interested in one dependent variable and designs a study to examine it by using an experiment or pure observation to see how the dependent variable changes under certain conditions (independent variables or factors). But the researcher is also aware that the dependent variable is likely to be influenced by other factors than the independent variables under study, namely covariates. In other words, the dependent variable covaries with these variables regardless of the changes (manipulated or not manipulated) in the independent variables. To control the effects of these covariates, analysis of covariance (ANCOVA) is used as the statistical method in data analysis.

A unique assumption for ANCOVA is homogeneity of group regression coefficients. The assumption states that within-group regression coefficients should be equal. In statistical terms, the assumption suggests that the regression slopes of all groups should be equal or parallel. When this assumption is violated, or evidence shows the slopes are not parallel, an interaction is present between the covariate (concomitant variable) and the independent variable (factor). In this case, interpreting that the changes in the dependent variable are influenced by the treatment/factor is invalid. Logically, such an interpretation should be avoided.

Homogeneity of group regression coefficients can be tested by contrasting two sources of variance: (a) deviation of the group regression coefficients from the



[a] Some theoretical statisticians (e.g., Lix et al., 1996; Wilcox, 1996, 1997) suggest when $p<.05$ from Levene's test, do not use conventional data analysis methods regardless of group conditions.
Note: s^2_L = the largest variance; s^2_S = the smallest variance; n_L = Largest group size; n_S = Smallest group size.

Figure 5 — A conceptual map for addressing assumption of equal variance.

average regression coefficient and (b) deviation of individual participants from their own group regression lines (Keppel, 1991). An *F* test is applied to the two sources of variance. When the *F* value from the test is greater than critical *F* value for the α (.10 or .25 as recommended by Lomax, 1998), then the assumption is violated. Keppel (1991) has provided a detailed description of the procedure for examining the assumption.

In addition, the data should satisfy the conventional assumptions, such as independence of observation, normality, and equal variance within/among groups. But ANCOVA is believed robust against violations of these assumptions as long as group sizes are equal. For violation of equal group regression coefficients, alternative statistical analysis methods should be applied. A desirable approach is Johnson-Neyman technique and Huitema (1980) has provided a detailed description of the technique and discussed the conditions for its optimal application.

MANOVA. MANOVA is used very often in physical education research. In MANOVA the equality of multiple means are compared. The three basic assumptions for inferential statistical analyses described above (independence of observation, normality of distribution, and equal variance) are extended to a set of assumptions in the multivariate analytical context. In this context, it is assumed that in each group (a) measures are taken from participants independently, (b) scores on all dependent variables follow a multivariate normal distribution, and (c) covariance for all dependent variables are equal. The independence assumption requires that each dependent variable should be measured independently. As described earlier, violation of this assumption can lead to serious inflation in Type I error (see Table 3).

The assumption of a multivariate normal distribution states that not only the scores of each dependent measure should be normally distributed, but also the joint distribution of scores of all dependent measures should be normal. When group sizes are equal and the assumption of equal covariances is met, all MANOVA procedures are rather robust against the effect of violation of the assumption on Type I error rate (Coombs, Algina, & Oltman, 1996; Stevens, 1992). In other words, when the assumption is violated, the chance of falsely rejecting the null hypotheses is relatively low; actual α usually deviates within a .02 range from the true α value. However, in severe violations (when nonnormality appears in all variables and in all groups), the statistical power of MANOVA to detect differences will be drastically reduced (Olson, 1974 as cited in Stevens, 1992). Thus, researchers may not be able to observe differences that do exist in the data. The procedure to test the assumption is presented by Stevens (1992, p. 247-256). Unfortunately, many popular statistical packages do not provide procedures for testing more than two dependent variables (bivariate normality).

The third assumption, equal covariances or homogeneity of variance—covariance matrices, states that the diagonal elements (representing the variances for the dependent variables) in one group's dependent variable matrix is equal to the corresponding elements in all other groups' matrices. MANOVA includes several procedures: Hotelling's T^2 for testing two-group mean vectors and multigroup testing procedures such as Roy's criterion, Pillai-Bartlett trace criterion, Wilks's criterion, and Hotelling-Lawley trace criterion. Violation of this assumption affects these procedures differently. For example, Hotelling's T^2 is not at all robust against the violation, especially when group sizes are unequal. When large

covariance is associated with the small group, the actual α will be liberal, leading to higher Type I error rates. When large covariance is associated with the large group, α will be conservative, resulting in weaker statistical power and a higher Type II error rate (Coombs et al., 1996).

Similar conclusions can be made for three of the four multigroup MANOVA procedures. Under both equal and unequal group size conditions, these procedures produce nonrobust actual α leading to a higher rate of Type I error. Olson's study (1976) suggests that Pillai-Bartlett's test is robust against the violation of equal covariance assumption under the condition of equal group size. In a replication research study, Elliott and Barcikowski (1994) confirmed Olson's conclusion by showing that Pillai-Bartlett's test maintains adequate statistical power under the data condition.

Alternatives to MANOVA procedures are available for researchers to use when the assumption of equal covariance is violated. Coombs et al. (1996) examined most alternatives and have provided recommendations for choosing among them. In studies involving two-group comparison, Hotelling's T^2 must be avoided when the assumption of equal covariance is violated, regardless of whether the group sizes are equal and distributions are normal.

As alternatives, Coombs et al. (1996) have recommended Johansen's test, James' first-order and second-order tests, Nel and van der Merwe procedure, Yao's test , and Kim's test. After evaluating the alternatives, Coombs et al. (1996) have suggested the following: (a) When the ratio of the smallest group size to the number of dependent variables is equal or greater than 4, Johansen's test is most adequate in replacing Hotelling's T^2; (b) when the ratio is smaller than 4, Kim's test may be considered; (c) when the multivariate normality assumption is violated, the above procedures will produce elevated Type I error rates. In this case, data transformations are needed to first satisfy the normality assumption (Stevens, 1992).

In multivariate studies involving multiple groups and/or variables (i.e., more than 3 groups and more than 4 dependent variables), violation of the equal covariance assumption seems to be inevitable (Coombs et al., 1996). When the assumption of multivariate normality is tenable, group sizes are adequately large ($n > 30$) and equal (ratio of largest to smallest < 1.5), the researcher may consider using Pillai-Bartlett trace criterion in MANOVA. When group sample sizes are small ($n < 20$) but equal, and the multivariate normality assumption is met, the researcher should consider using alternative analysis methods. As recommended by Coombs et al. (1996), Johansen's and James' second-order tests will perform well in controlling Type I error rates. When the group sizes are unequal, but the multivariate normality assumption is met, Johansen's test may be the first choice. An additional data condition required for Johansen's test is that the ratio of smallest group size to the number of dependent variables is greater than 3 1/3 when there are three independent groups, or greater than 4 2/3 when there are more groups. James' second-order test is recommended when the ratio is greater than 4.

All the alternatives are based on the assumption of multivariate normality. When this assumption and the assumption of equal covariance matrices are violated simultaneously, the researcher should seek appropriate data transformations. Stevens (1992, p. 251-256) provides a detailed rationale and graphical explanations for data transformations. Again, interpreting results from transformed data is always challenging. Thus, transformation may not be an ideal way to handle the problem.

The latest developments in theoretical statistics have shown promising analysis methods to control Type I error rates without sacrificing adequate statistical power. Wilcox (1997) has summarized these approaches including the Winsorizing and trimming technique, which is one that gives different weights to scores located at different distribution places (e.g., center or tail). A Winsorized distribution, then, is a function of the true distribution weighted more toward the center of the true distribution. Trimming, on the other hand, is simply to remove the tails of a distribution (Wilcox, 1997).

Conclusions

When answering research questions involves hypothesis testing, inferential statistics may be the best tool for achieving "conclusion coherence" (Levin & Robinson, 2000). Conclusion coherence means that interpretations of the results are solely based on valid data and analyses. In other words, results of a study should be based on a "real" effect rather than "chance." Thus, strictly following the requirements for examining data conditions and assumptions is a critical step for achieving conclusion coherence.

Information presented in this article shows that conducting statistical analyses without considering the assumptions is inappropriate. Examining the assumptions in data analysis seems to have been overlooked in pedagogy research in physical education. In addition, the conventional wisdom about the robustness of inferential statistical analyses is being challenged in current statistical literature. It is recommended that remedial procedures should be adopted in order to generate valid results from statistical analyses.

Although the publication manual of the American Psychology Association (1994) and many journal editorial guidelines do not require researchers to report results of assumption examinations, scholars in research methodology often consider it necessary to include them in the research report. Huck (2000), for example, has recommended that researchers should include in their reports brief statements describing results of assumption tests, efforts to meet assumptions, and a rationale for choosing alternative analyses when the assumptions are violated. Huck argues whether reporting this information is a criterion to evaluate the quality of a research study. As can be seen in Table 2, only a fraction of physical education research articles has presented information that meets this challenge.

In this paper, we have shown the importance of meeting the assumptions of inferential statistical analysis and possible consequences of their violations. Although it is seldom that data from school-based research can meet the assumptions perfectly, it is crucial that the researcher understands the condition of the data and makes necessary adjustment in the process of data analysis to preserve validity of the results. To conclude, we draw upon Robert W. Schutz's comment on the importance of knowing data condition in data analysis. In a short article for a theme column of the *Measurement News*, "The most important things that I have learned (so far)," Schutz summarized what he has learned throughout his distinguished career as a measurement and statistics scholar. He wrote: " . . . if the question . . . is 'what should they [researchers] know?' then I suggest the following (which I learned quite early in my career as being extremely important) – KNOW YOUR DATA" (Schutz, 2001, p. 1; parenthesis and emphasis original; brackets added).

References

Alexander, R.A., & Govern, D.M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics*, **19**, 91-101.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

Bishop, T.A., & Dudewicz, E.J. (1978). Exact analysis of variance with unequal variances: Test procedures and tables. *Technometrics*, **20**, 419-430.

Bock, R.D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Borg, W.R., & Gall, M.D. (1989). *Educational research: An introduction*. White Plains, NY: Longman.

Box, G.E.P. (1954). Some thorems on quardratic forms applied in the study of analysis of variance problems, I. Effect of inqueality of variance in the one-way classification. *Annals of Mathematical Statistics*, **25**, 290-302.

Bradley, J.W. (1980). Nonrobustness in z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society*, **16**, 333-336.

Brown, M.B., & Forsythe, A.B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association*, **69**, 364-367.

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

Clinch, J.J., & Keselman, H.J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics*, **7**, 207-214.

Conover, W.J., Johnson, M.E., & Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, **23**, 351-361.

Coombs, W.T., Algina, J., & Oltman, D.O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal. *Review of Educational Research*, **66**, 137-179.

Elliott, R.S., & Barcikowski, R.S. (1994). Investigation of power using F approximations for the Hotelling-Lawley trace and Pillai's trace. *Mid-Western Educational Researcher*, **7**, 2-6.

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

Heck, R.H., & Thomas, S.L. (2000). *An introduction to multilevel modeling technique*. Mahwah, NJ: LEA.

Hjorth, U. (1994). *Computer intensive statistical methods: Validation model selection and bootstrap*. New York: Chapman & Hall.

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

Huitema, B.E. (1980). *The analysis of covariance and alternatives*. New York: Wiley.

Huitema, B.E., McKean, J.W., & McKnight, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement*, **59**, 767-786.

James, G.S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometricka*, **38**, 324-329.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

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.

Kirk, R.E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

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

Kruskal, W.H., & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**, 583-621.

Lee, V.E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, **35**, 125-141.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278-292). Stanford, CA: Stanford University Press.

Levin, J.R., & Robinson, D.H. (2000). Rejoinder: Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher*, **34**(1), 34-36.

Lix, L.M., & Keselman, H.J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, **58**, 409-429.

Lix, L.M., Keselman, J.C., & Keselman, H.J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance *F* test. *Review of Educational Research*, **66**, 579-619.

Lomax, R.G. (1998). *Statistical concepts: A second course for education and the behavioral sciences*. Mahwah, NJ: LEA.

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

Noreen, W.E. (1989). *Computer intensive methods for testing hypotheses: An introduction*. New York: Wiley-Interscience.

Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, **69**, 894-908.

Olson, C.L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, **83**, 579-586.

Oshima, T.C., & Algina, J. (1992). Type I error rates for James' second-order test and Wilcox's H_m test under heteroscedasticity and non-normality. *British Journal of Mathematical and Statistical Psychology*, **42**, 255-263.

Pearson, E.S., & Please, N.W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, **62**, 223-241.

Scariano, S., & Davenport, J. (1987). The effects of violations of independence assumptions in the one-way ANOVA. *The American Statistician*, **41**, 123-129.

Schutz, R.W. (2001). The most important things that I have learned (so far). *Measurement News*, **6**(1), 1.

Silverman, S., & Solmon, M. (1998). The unit of analysis in field research: Issues and approaches to design and data analysis. *Journal of Teaching in Physical Education*, **17**, 270-284.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Stevens, J. (1990). *Intermediate statistics: A modern approach*. Hillsdale, NJ: LEA.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: LEA.

Tan, W.Y. (1982). Sampling distributions and robustness of *t*, *F* and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communications in Statistics*, **A11**, 2485-2511.

Thomas, J.R., Nelson, J.K., & Thomas, K.T. (1999). A generalized rank-order method for nonparametric analysis of data from exercise science: A tutorial. *Research Quarterly for Exercise and Sport*, **70**, 11-23.

Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, **38**, 330-336.

Wilcox, R.R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, **38**, 29-60.

Wilcox, R.R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, **65**, 51-77.

Wilcox, R.R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilcox, R.R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

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

Yuen, K.K. (1974). The two-sample trimmed t for unequal population variance. *Biometrika*, **61**, 165-170.

Zhu, W. (1997a). A multilevel analysis of school factors associated with health-related fitness. *Research Quarterly for Exercise and Sport*, **68**, 125-135.

Zhu, W. (1997b). Making bootstrap statistical inferences: A tutorial. *Research Quarterly for Exercise and Sport*, **68**, 44-55.