Modern Robust Statistical Methods

An Easy Way to Maximize the Accuracy and Power of Your Research

David M. Erceg-Hurn — *University of Western Australia*
Vikki M. Mirosevich — *Government of Western Australia*

Classic parametric statistical significance tests, such as analysis of variance and least squares regression, are widely used by researchers in many disciplines, including psychology. For classic parametric tests to produce accurate results, the assumptions underlying them (e.g., normality and homoscedasticity) must be satisfied. These assumptions are rarely met when analyzing real data. The use of classic parametric methods with violated assumptions can result in the inaccurate computation of p *values, effect sizes, and confidence intervals. This may lead to substantive errors in the interpretation of data. Many modern robust statistical methods alleviate the problems inherent in using parametric methods with violated assumptions, yet modern methods are rarely used by researchers. The authors examine why this is the case, arguing that most researchers are unaware of the serious limitations of classic methods and are unfamiliar with modern alternatives. A range of modern robust and rank-based significance tests suitable for analyzing a wide range of designs is introduced. Practical advice on conducting modern analyses using software such as SPSS, SAS, and R is provided. The authors conclude by discussing robust effect size indices.*

Keywords: robust statistics, nonparametric statistics, effect size, significance testing, software

Null hypothesis significance testing is the workhorse of research in many disciplines, including medicine, education, ecology, economics, sociology, and psychology. A recent study of 10 leading international psychology journals found that null hypothesis significance testing was used in 97% of articles (Cumming et al., 2007). The most widely used null hypothesis tests are classic parametric procedures, such as Student's *t*, analysis of variance (ANOVA), and ordinary least squares regression. For classic parametric tests to produce accurate results, the assumptions underlying them must be sufficiently satisfied. However, these assumptions are rarely met when analyzing real data. The use of classic parametric tests when assumptions are violated can lead to the inaccurate calculation of *p* values. This can result in an increased risk of falsely rejecting the null hypothesis (i.e., concluding that real effects exist when they do not). In contrast, power to detect genuine effects is often substantially reduced. An additional problem is that common measures of effect size (e.g., Cohen's *d*) and confidence intervals may be inaccurately estimated when classic parametric assumptions (e.g., normality) are violated. The miscomputation of *p* values, coupled with the inaccurate estimation of effect sizes and confidence intervals, can lead to substantive errors in the interpretation of data. Several prominent statisticians and researchers have described the use of classic parametric statistics in the face of assumption violations as invalid (e.g., Kezelman et al., 1998; Leech & Onwuegbuzie, 2002; Wilcox, 2001; Zimmerman, 1998).

Modern robust statistical procedures exist that can solve the problems inherent in using classic parametric methods when assumptions are violated. Many modern statistical procedures are easily conducted with widely used software such as SPSS, SAS, and R. Despite the advantages of modern methods and the ease with which these procedures can be conducted, they are rarely used by researchers. In the first part of this article, we examine why this is the case. We argue that most researchers are unaware of the limitations of classic methods and do not realize that modern alternatives exist. In the second half of the article, we provide a practical, nontechnical introduction to some modern methods.

Problems With Classic Parametric Methods

Classic parametric methods are based on certain assumptions. One important assumption is that the data being analyzed are normally distributed. In practice, this assumption is rarely met. Micceri (1989) examined 440 large data sets from the psychological and educational literature, including a wide range of ability and aptitude measures (e.g., math and reading tests) and psychometric measures (e.g., scales measuring personality, anxiety, anger, satisfaction, locus of control). None of the data were normally distributed, and few distributions remotely resembled the normal curve. Instead, the distributions were frequently multimodal, skewed, and heavy tailed. Micceri's study indicated that real data are more likely to resemble an exponential

David M. Erceg-Hurn, School of Psychology, University of Western Australia, Crawley, Western Australia, Australia; Vikki M. Mirosevich, Department of Health, Government of Western Australia.

We thank Kale Dyer and Rand Wilcox for their helpful feedback on drafts of this article. Thanks also to Amy Lampard and Jake England for their support and encouragement.

Correspondence concerning this article should be addressed to David M. Erceg-Hurn, School of Psychology, M304, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia. E-mail: david.erceg-hurn@grs.uwa.edu.au

 0003-066X/08/$12.00
Vol. 63, No. 7, 591–601 DOI: 10.1037/0003-066X.63.7.591

David M. Erceg-Hurn

curve than a normal distribution. Micceri's findings are consistent with other research. Bradley (1977) identified several examples of asymmetrical and skewed distributions in the social science literature. For example, reaction time is often used as a dependent variable in psychological research, and it is well-known that reaction time data are frequently skewed (Miller, 1988; Taylor, 1965).

Another important assumption underlying classic parametric tests is that of equal population variances (also called homogeneity of variance, or homoscedasticity). The assumption can be framed in terms of a variance ratio (VR). If two populations have similar variances, their VR will be close to 1:1. For example, if the variance of Population A is 120 and the variance of Population B is 100, their VR would be 120:100, or 1.2:1. When real data are analyzed, the VR often strays markedly from the 1:1 ratio required to fulfill the assumption. Keselman et al. (1998) conducted a comprehensive review of ANOVA analyses in 17 educational and child psychology journals. For each study, a sample VR was calculated by dividing the largest variance by the smallest variance. In studies using a one-way design, the median VR was 2.25:1, and the mean VR was 4:1. In factorial studies, the median VR was 2.89:1 and the mean VR was 7.84:1. Keselman et al. identified several extreme VRs, the largest being 566:1. Large variance ratios have also been found in reviews of studies published in clinical and experimental psychology journals. Grissom (2000) examined one issue of the *Journal of Consulting and Clinical Psychology* and identified sample ratios of 4:1, 6:1, 7:1, and 8:1 on more than one occasion. Ratios of 25:1 and 281:1 were also found. We examined two recent issues of the *Journal of Experimental Psychology: General* and the *Journal of Experimental Psychology: Human Perception and Performance*. We identified 28 studies in which data were analyzed using ANOVA. In these studies, sample VRs between 2:1 and 4:1 were common. Several VRs exceeding 10:1 were identified, including VRs of 39:1, 59:1, 69:1, and 121:1. These sample VRs are subject to considerable sampling error; nevertheless, the magnitude of the VRs and the consistency with which they are reported suggest that it is not unusual for the homoscedasticity assumption to be violated.

The presence of heteroscedasticity in real data is not surprising, given the nature of the research designs and samples psychologists typically use. Researchers are often interested in comparing the performance of preexisting groups (e.g., men and women) on some dependent variable. Groups defined using a preexisting factor can have different variances (Keppel & Wickens, 2004). For example, the performance of older adults on measures of cognitive functioning is more variable than that of younger adults (Hultsch, MacDonald, & Dixon, 2002). Heteroscedasticity can also occur in completely randomized experiments as a result of the experimental variable causing differences in variability between groups. Consider a trial investigating the efficacy of a novel psychotherapy for depression. Participants in the trial are randomly allocated to either (a) an experimental group that receives the novel treatment or (b) a control group that receives no treatment. Participants' depressive symptoms are measured at the start of the trial and 12 weeks later. Because of random allocation, the variances of the two groups at the start of the trial should be roughly equivalent. However, the groups' variances at the end of the trial may be significantly different as a result of the effects of the experimental variable. There may be great variability in the response of participants in the novel psychotherapy group. Some participants may find that their symptoms completely remit, others may have a partial response, some may experience no change, and a few may experience worsening of their symptoms. In contrast, the majority of the participants in the control group may experience comparatively little change in their depressive symptoms. If this is the case, the variances of the two groups at the end of the trial will be heterogeneous. See Grissom and Kim (2005) for additional discussion about why heteroscedasticity occurs in real data.

Violation of the normality and homoscedasticity assumptions can have a substantial influence on the results of classic parametric tests, in particular on rates of Type I and Type II error. A Type I error occurs when the null hypothesis is falsely rejected. In other words, one concludes that a real effect exists when it does not. In contrast, a Type II error occurs when the null hypothesis is not rejected even though it is false. The power of a test is the probability that a Type II error will not occur.

Violation of the normality and homoscedasticity assumptions can cause the Type I error rate to distort. Usually, the Type I error rate (also known as the alpha rate, or α) is set at .05. This means that if a result is deemed statistically significant, there should be less than a 5% risk that a Type I error has been made. However, when classic parametric tests are used to analyze nonnormal or heteroscedastic data, the true risk of making a Type I error

Vikki M. Mirosevich

may be much higher (or lower) than the obtained *p* value. Consider performing an ANOVA on data when variances and sample sizes are unequal. An ANOVA is run in SPSS or SAS, and the *p* value reported is .05. This should mean that if the null hypothesis is rejected, there is less than a 5% chance that a Type I error has been made. However, the true risk of committing a Type I error may be closer to 30% (Wilcox, Charlin, & Thompson, 1986). Contrary to popular belief, equal sample sizes offer little protection against inflated Type I error when variances are heterogeneous (Harwell, Rubinstein, Hayes, & Olds, 1992). Conducting a regression analysis with violated assumptions can also lead to inflated rates of Type I error. The probability of a Type I error when testing at an alpha rate of .05 can exceed 50% when data are nonnormal and heteroscedastic (Wilcox, 2003). Researchers need to be aware that the *p* values reported by statistical packages such as SPSS may be extremely inaccurate if the data being analyzed are nonnormal and/or heteroscedastic; The inaccuracy may lead researchers to unwittingly make Type I errors.

An additional problem is that the power of classic parametric tests can be substantially lowered when the assumptions of normality or homoscedasticity are violated. See Wilcox (1998) for an example in which only a small departure from normality reduces the power of the *t* test from .96 to .28. Wilcox (1998) summarized the effect on power of violating the normality and homoscedasticity assumptions as follows:

> As hundreds of articles in statistical journals have pointed out and for reasons summarized in several books . . . arbitrarily small departures from normality can result in low power; even when distributions are normal, heteroscedasticity can seriously lower the power of standard ANOVA and regression methods. The practical result is that in applied work, many nonsignificant results would have been significant if a more modern method, developed after the year 1960, had been used. (p. 300)

As noted by Wilcox (1998), modern robust statistics exist that can solve many of the problems caused by violating the assumptions of classic parametric tests. The term *robust statistics* refers to procedures that are able to maintain the Type I error rate of a test at its nominal level and also maintain the power of the test, even when data are nonnormal and heteroscedastic (see Wilcox, 2005, for a more detailed discussion of statistical criteria for judging robustness). Countless studies have shown that, in terms of Type I error control and statistical power, modern robust statistics frequently offer significant advantages over classic parametric methods, particularly when data are not normally distributed or heteroscedastic. However, modern robust methods are rarely used by researchers. We now examine why most researchers do not make use of the wide array of robust statistics that have been developed over the past 50 years.

Why Are Modern Methods Underused?

Lack of Familiarity With Modern Methods

Most researchers do not realize that modern robust statistical methods exist. This is largely due to lack of exposure to these methods. For example, the psychology statistics curriculum, journal articles, popular textbooks, and software are dominated by statistics developed before the 1960s. This problem is not limited to psychology but also exists in many other disciplines (e.g., medicine, ecology). The field of statistics has progressed markedly since 1960, yet most researchers rely on outdated methods. We are not trying to blame researchers for not being familiar with modern statistical methods. Researchers are busy in their own areas of expertise and cannot be expected to be familiar with cutting-edge developments within statistics. At the same time, it is essential that researchers are made aware of important developments within statistics that have the potential to improve research in domains such as psychology.

Assumption Testing Issues

Another reason why modern methods are underused is that researchers frequently fail to check whether the data they are analyzing meet the assumptions underlying classic parametric tests (Keselman et al., 1998). This may be due to forgetfulness or not knowing how to check assumptions. A related problem is that, due to low power, statistical assumption tests built into software such as SPSS often do a poor job of detecting violations from normality and homoscedasticity (Jaccard & Guilamo-Ramos, 2002). For example, Levene's test is often used to test the homoscedasticity assumption. A *p* value greater than .05 is usually taken as evidence that the assumption has not been violated. However, Levene's test can yield a *p* value greater than .05, even when variances are unequal to a degree that could significantly affect the results of a classic parametric test. This is particularly true when small samples are being

analyzed. Another problem is that assumption tests have their own assumptions. Normality tests usually assume that data are homoscedastic; tests of homoscedasticity assume that data are normally distributed. If the normality and homoscedasticity assumptions are violated, the validity of the assumption tests can be seriously compromised. Prominent statisticians have described the assumption tests (e.g., Levene's test, the Kolmogorov–Smirnov test) built into software such as SPSS as fatally flawed and recommended that these tests never be used (D'Agostino, 1986; Glass & Hopkins, 1996). The take-home message is that researchers should not rely on statistical tests to check assumptions because of the frequency with which they produce inaccurate results.

The Robustness Argument

Researchers often claim that classic parametric tests are robust to violations of the assumptions of normality and homoscedasticity, negating the need to use alternative procedures. *Robust* in this sense is generally taken to mean that the tests maintain rates of Type I error close to the nominal level. Note the difference between this definition of *robust* and the definition of *robust statistics* given earlier. Robust statistics control Type I error and also maintain adequate statistical power. In contrast, claims that classic parametric tests are robust usually only consider Type I error, not power. An overview of the robustness argument can be found in Sawilowsky (1990).

The origin of the robustness argument can be traced back to several key articles and books, including Boneau (1960); Box (1953); Lindquist (1953); and Glass, Peckham, and Sanders (1972). These widely cited studies concluded that classic parametric methods are exceedingly robust to assumption violations. These claims of robustness found their way into introductory statistical textbooks, and researchers quickly came to accept as fact the notion that classic parametric tests are robust. See Bradley (1978) for an excellent summary of how this occurred. Today, the belief that classic parametric tests are robust is widespread (Wilcox, 1998). Most research methods textbooks used by researchers continue to claim that classic tests are generally robust, at least for balanced designs. However, there are several problems with the robustness argument. The early studies cited previously only examined the impact of small deviations from normality and homoscedasticity, not the large deviations that are often found when analyzing real psychological data. Therefore, the early studies do not provide a valid assessment of how classic parametric tests perform under real-world data analytic conditions. Also, the studies generally investigated the impact of violating normality and homoscedasticity in isolation, whereas in practice it is often the case that both assumptions are concurrently violated (Bradley, 1980; Keppel & Wickens, 2004). Furthermore, several authors (e.g., Bradley, 1978; Harwell, 1992) have noted that a careful reading of the early studies allows for very different conclusions about robustness to be reached. Bradley pointed out that the authors of the early studies downplayed evidence that did not support their arguments and overextended assertions of robustness beyond their data, claiming that classic parametric tests are robust in a wide range of circumstances.

Considerable research indicates that classic parametric tests are only robust in a limited number of circumstances, not the vast majority as is widely believed. For example, Sawilowsky and Blair (1992) found that the *t* test is relatively robust to violation of the normality assumption when the following four conditions hold: (a) variances are equal, (b) sample sizes are equal, (c) sample sizes are 25 or more per group, and (d) tests are two-tailed. This combination of conditions is not reflective of most real data analytic circumstances, where unequal sample sizes are common and variances are often heterogeneous. Sawilowsky and Blair found that when one-tailed tests were used, the Type I error rate would become conservative. Several researchers have shown that the *t* test is not robust when the homogeneity of variance assumption is violated, nor is it robust when the normality and homogeneity of variance assumptions are concurrently violated (e.g., Ramsey, 1980; Zimmerman, 1998). In most situations—particularly when analyzing real-world data—robustness is the exception rather than the rule.

Proponents of the robustness argument have typically focused their attention on Type I error but have not considered the power of classic parametric tests when data are nonnormal or heteroscedastic. Countless studies have shown that even when classic parametric tests are robust to Type I errors, they are usually considerably less powerful than their modern robust counterparts. For example, Akritas, Arnold, and Brunner (1997) demonstrated that when the normality assumption is violated, a modern version of ANOVA can be more than three times as powerful as the classic ANOVA used by most researchers. Even if researchers insist that classic parametric tests are robust, this does not preclude the use of alternate procedures. Modern methods are also robust and more powerful when data are not normally distributed and/or heteroscedastic.

Transformations

Rather than using modern methods, researchers sometimes opt to transform their data. In these cases, a transformation such as the square root or logarithm is performed, and classic parametric tests are used to analyze the transformed data. The use of transformations is problematic for numerous reasons, including (a) transformations often fail to restore normality and homoscedasticity; (b) they do not deal with outliers; (c) they can reduce power; (d) they sometimes rearrange the order of the means from what they were originally; and (e) they make the interpretation of results difficult, as findings are based on the transformed rather than the original data (Grissom, 2000; Leech & Onwuegbuzie, 2002; Lix, Keselman, & Keselman, 1996). We strongly recommend using modern robust methods instead of conducting classic parametric analyses on transformed data.

Classic Nonparametric Statistics

Several classic nonparametric statistics are built into widely used software such as SPSS and SAS. Some researchers elect to use these classic nonparametric statistics

rather than modern methods. As with classic parametric techniques, classic nonparametric tests were developed before the 1960s and suffer from many limitations. For example, classic nonparametric statistics are not robust when used to analyze heteroscedastic data (Harwell et al., 1992; Lix et al., 1996; Sawilowsky, 1990; Zimmerman, 1998, 2000). Another major limitation is that classic nonparametric tests are only appropriate for analyzing simple, one-way layouts and not factorial designs involving interactions. Modern robust methods (which include modern nonparametric procedures) are not susceptible to these limitations.

Misconceptions About Modern Methods

Some researchers have misconceptions about modern methods that have contributed to the underuse of these procedures. One misconception is that software to perform modern statistical analyses is not readily available. This belief may stem from the fact that modern robust statistics are not built into widely used statistical software such as SPSS and SAS. This has made modern methods invisible to many researchers. Fortunately, proponents of modern methods have created special software add-ons that allow researchers to conduct analyses using SPSS and SAS. Furthermore, a vast array of alternative, free software is available that can conduct modern analyses.

Another misconception held by some researchers is that modern methods should not be used because they sometimes involve trimming or ranking procedures that discard valuable information. Wilcox (2001) noted that it is somewhat counterintuitive that a test could be more accurate by removing information—hence why some researchers are suspicious of modern methods. However, take the cases of outliers (highly unusual data points) and trimmed means. Consider a data set containing the following values:

1, 1, 1, 2, 2, 5, 5, 5, 6, 20, 40.

The mean of the values is 8. However, the mean is distorted by two outlying values (20 and 40). All of the other values in the data set are less than or equal to 6. Consequently, the mean does not accurately reflect the central values of the data set. Instead of using the mean as a measure of central tendency, we could instead use the median, which in this case is 5. The median is an extreme form of a trimmed mean, in the sense that all but the middle score is trimmed. However, calculating the median discards a lot of information, as every value above and below the middle point of the data set is removed. A compromise between the mean and the median is the 20% trimmed mean. To obtain the 20% trimmed mean, we remove the lowest and highest 20% of the values from the data set, leaving

1, 2, 2, 5, 5, 5, 6.

The mean of the remaining values is then calculated. In this case, the 20% trimmed mean is 3.71, which reflects the central values of the original data set more accurately than the untrimmed mean of 8. The trimmed mean is an attractive alternative to the mean and the median, because it effectively deals with outliers without discarding most of the information in the data set. Research has shown that the use of trimming (and other modern procedures) results in substantial gains in terms of control of Type I error, power, and narrowing confidence intervals (Keselman, Algina, Lix, Wilcox, & Deering, 2008; Wilcox, 2001, 2003, 2005). Furthermore, if data are normally distributed, the mean and the trimmed mean will be the same.

A Practical Introduction to Modern Methods

What follows is a practical, nontechnical introduction to some modern robust statistical methods. Modern robust methods are designed to perform well when classic assumptions are met, as well as when they are violated. Therefore, researchers have little to lose and much to gain by routinely using modern statistical methods instead of classical techniques. An alternative strategy is to analyze data using both classic and modern methods. If both analyses lead to the same substantive interpretation of the data, debate about which analysis should be trusted is moot. If classic and modern analyses lead to conflicting interpretations of data, the reason for the discrepancy should be investigated. Differences will often be due to nonnormality, heteroscedasticity, or outliers causing classic techniques to produce erroneous results. Consequently, analyses conducted using modern methods should usually be trusted over those conducted using classic procedures. However, each situation needs to be assessed on its own merits. Because of the serious limitations of assumption tests noted earlier, researchers should not use assumption tests as a basis for deciding whether to use classic or modern statistical techniques.

The defining feature of robust statistics is that they are able to maintain adequate Type I error control and statistical power, even when data are nonnormal or heteroscedastic. Essentially, robust methods work by replacing traditional regression estimators (i.e., ordinary least squares), measures of location (e.g., the mean) and measures of association (e.g., Pearson's *r*) with robust alternatives. Hypothesis testing can be performed using these robust measures. For example, Keselman et al. (2008) proposed a robust approach to hypothesis testing that involves trimmed means, Winsorized variances, and bootstrapping. Keselman et al. recommend the use of 20% trimmed means, although on some occasions a smaller or larger amount of trimming may be desirable (see Wilcox, 2005, p. 57).

Winsorized Variance

Keselman et al.'s (2008) robust approach to hypothesis testing involves the replacement of a distribution's variance with a robust alternative, the Winsorized variance. The benefit of the Winsorized variance is that it is more resistant to outliers than the variance is. The use of Winsorizing can result in the estimation of more accurate standard errors than if classic methods are used.

To understand the calculation of a Winsorized variance, imagine that a study is conducted with 10 participants, who have the following scores on the dependent variable:

3, 1, 75, 10, 5, 6, 11, 7, 75, 12.

The first step in computing the Winsorized variance is to reorder the scores from lowest to highest:

1, 3, 5, 6, 7, 10, 11, 12, 75, 75.

The second step in Winsorizing (if 20% trimming is being used) is to remove the lowest and highest 20% of scores from the data set. In this case, the scores 1, 3, 75, and 75 will be removed, leaving

5, 6, 7, 10, 11, 12.

Next, the removed scores in the lower tail of the distribution are replaced by the smallest untrimmed score, and the removed scores in the upper tail of the distribution are replaced by the highest untrimmed score. The untrimmed and replaced scores are known as *Winsorized scores*. For our data set, the Winsorized scores are

5, 5, 5, 6, 7, 10, 11, 12, 12, 12.

The mean of the Winsorized scores is then calculated:

$$\bar{X}_w = \frac{1}{10}(5 + 5 + 5 + 6 + 7 + 10 + 11 + 12 + 12 + 12) = 8.5. \quad (1)$$

Finally, the variance of the Winsorized scores is calculated by using the same formula that is used to calculate the (ordinary least squares) variance, except that the Winsorized scores and Winsorized mean are used in place of the original scores and mean. Therefore, any software program that can calculate variance can also be used to calculate Winsorized variance. For the present data set, the Winsorized variance is 90.5.

Bootstrapping

Bootstrapping is a computer-intensive resampling technique. All bootstrap methods involve generating bootstrap samples based on the original observations in a study. Consider a study in which the following scores on the dependent variable are observed:

2, 3, 3, 4, 5, 6, 7, 8, 9, 9, 9, 10.

The sample size is 12, and the sample mean is 6.25. In calculating a bootstrap sample, a computer is used to randomly sample with replacement 12 observations one at a time from the original scores. *Sampling with replacement* means that each individual score remains in the original data set before the selection of the next score rather than being removed from the original data set. As a result, observations can occur more (or fewer) times in the bootstrapped sample than they did in the original sample. A bootstrap sample generated from the original observations in this example might be

3, 3, 3, 3, 4, 4, 7, 8, 8, 9, 10, 10.

The mean of this bootstrap sample is 6. The process of generating bootstrap samples from the original scores is repeated hundreds or thousands of times. With modern computers, this can be accomplished in seconds.

Bootstrapping is often used to get a better approximation of the sampling distribution of a statistic (e.g., the *t* distribution) than its theoretical distribution provides when assumptions are violated. In other words, instead of assuming that the data collected follow the *t*, chi-square, or some other distribution, bootstrapping is used to create a sampling distribution, and this bootstrapped distribution can be used to compute *p* values and test hypotheses. Bootstrapping can also be used to generate confidence intervals. For example, imagine that we want to create a 95% confidence interval around a mean. We could accomplish this using the percentile bootstrap method. Imagine that a study is conducted and that the mean of participants on the dependent variable is 6.50. We use the participants' scores to generate 1,000 bootstrap samples. For each bootstrap sample, a mean is calculated. The 1,000 bootstrapped means are then put in order, from lowest to highest, and the central 95% of values are used to form the confidence interval. If the central 95% of values fall between 4.70 and 7.80, these values would form the lower and upper limits of the 95% confidence interval around the mean. Software is available that will conduct analyses based on bootstrap samples (see below).

Robust Hypothesis Testing, Software, and Resources

The robust approach to hypothesis testing proposed by Keselman et al. (2008) uses trimmed means, Winsorized variances, and bootstrapping to calculate a test statistic and *p* value, which they term the *adjusted degrees of freedom* (ADF) solution. The ADF solution can be used to evaluate hypotheses analogous to those tested using classic parametric tests. The only difference is that the hypotheses evaluated using Keselman et al.'s approach concern trimmed means rather than means. For example, the null hypothesis tested using the classic independent groups *t* test is that two population means are equal, whereas in Keselman et al.'s approach, the null hypothesis is that two population-trimmed means are equal. Keselman et al. have developed a free SAS/IML program that can be used to perform hypothesis testing using the ADF solution. The program can also compute robust estimates of effect size. The program and instructions are available from the American Psychological Association Web site, http://dx.doi.org/10.1037/1082-989X.13.2.110.supp. The instructions outline how the SAS program analyzes data for one-way and factorial designs.

The ADF solution is just one of many robust hypothesis testing methods developed over the past 40 years. Readers interested in a short introduction to some modern methods may benefit from consulting Wilcox and Keselman (2003). There are also several excellent books about robust methods that researchers will find useful. Wilcox (2001) is an eminently readable, nontechnical introduction to the fundamentals of robust statistics. For a

clear, comprehensive, and practical overview of a wide variety of robust methods, see Wilcox (2003). More advanced coverage is found in Wilcox (2005). Wilcox has written code for R to conduct the analyses described in his 2003 and 2005 texts. R is a powerful statistical software package, free to download from http://www.R-project.org. R uses a command line (similar to SPSS's syntax) rather than a graphical user interface. Wilcox's code can estimate numerous measures of location and scale; detect outliers; calculate confidence intervals; analyze data from one-way, factorial, repeated measures, split plot, and multivariate designs; and perform multiple comparison procedures, correlations, and robust regression. The latest version of Wilcox's code is available as a free download from http://www-rcf.usc.edu/~rwilcox.

Researchers who prefer to work with a graphical user interface may be interested in ZumaStat, an easy-to-use and relatively inexpensive software package developed by James Jaccard. ZumaStat allows users to conduct robust analyses from within SPSS or Microsoft Excel. ZumaStat adds a toolbar to SPSS or Excel that allows users to point and click to select robust statistical procedures. The software then feeds instructions into R, which performs the relevant analyses. The ZumaStat software can perform most of Wilcox's R functions. For further information, visit http://www.zumastat.com.

SAS/STAT 9 includes some inbuilt capability for performing robust analyses. The ROBUSTREG procedure can perform a limited number of robust regression techniques, such as M estimation and least trimmed squares. These procedures can also be used for ANOVA (SAS Institute, 2004, pp. 3971–4030).

Modern Rank Statistics

The modern robust methods discussed thus far can be thought of as modern day versions of classic parametric procedures, such as ANOVA and least squares regression. These techniques revolve around robust measures of location and scale, such as trimmed means and Winsorized variances. Another set of modern techniques have been developed that revolve around ranked data. These rank-based techniques can be thought of as modern versions (and extensions) of classic nonparametric statistics. Modern rank-based procedures are robust in the sense that they produce valid results when analyzing data that is nonnormal and/or heteroscedastic. We now briefly introduce some prominent modern rank-based methods.

Rank Transform

Conover and Iman (1981) proposed a simple, two-step procedure known as the rank transform (RT). RT is only mentioned here so that researchers know to avoid it. The RT procedure involves (a) converting data to ranks and (b) performing a standard parametric analysis on the ranked data instead of original scores. The appeal of RT is that it is easily conducted using software such as SPSS and SAS. In fact, official SAS user documentation encourages the use of RT (SAS Institute, 2004). Early research into the technique was promising; however, by the late 1980s, numerous problems with RT had emerged. See Fahoome and Sawilowsky (2000), Lix et al. (1996), Sawilowsky (1990), and Toothaker and Newman (1994) for concise reviews of this research. RT can perform well, but in many circumstances it is nonrobust and can be less powerful than classic parametric and nonparametric methods. The consensus in the literature is that RT should not be used.

ANOVA-Type Statistic

An alternative to the RT is the ANOVA-type statistic (ATS; Brunner, Domhof, & Langer, 2002; Brunner & Puri, 2001; Shah & Madden, 2004). As its name suggests, the ATS can be used in data analytic situations in which ANOVA is traditionally used. The hypotheses tested by ANOVA and the ATS are similar. An ANOVA assumes that the groups being compared have identical distributions and tests the null hypothesis that the means of the groups are the same. The ATS tests the null hypothesis that the groups being compared have identical distributions and that their relative treatment effects ($\hat{p}_i$) are the same. A relative treatment effect is the tendency for participants in one group to have higher (or lower) scores on the dependent variable, compared with the scores of all participants in a study. Relative treatment effects can range between 0 and 1 (if the null hypothesis is true, all groups should have a relative treatment effect of .50).

To understand the computation and interpretation of relative treatment effects, consider a study in which a researcher is interested in comparing three groups (A, B, C) on some dependent variable. There are four participants in each group. Their scores and the following calculations are shown in Table 1. First, convert the scores to ranks, ignoring group membership. That is, the smallest score in the entire data set is assigned a rank of 1, the second smallest score a rank of 2, and so on until all scores have been converted to ranks. Tied scores are assigned midranks (i.e., the second and third score in this data set are both 11, so the assigned ranks would be 2 and 3, but they are given an average rank, or midrank, of $(2 + 3)/2 = 2.5$. Next, the ranks in each group are summed and then divided by the number of observations in the group to calculate each group's mean rank. For example, the sum of the ranks in Group A is 21.5, and there are four observations. Therefore, the mean rank of Group A is simply $21.5/4 = 5.375$. Finally, the relative treatment effect of each group is computed using the following equation:

$$\hat{p}_i = \frac{\bar{R}_{i\bullet} - .50}{N}, \tag{2}$$

where $\hat{p}_i$ is the ith group's estimated relative treatment effect, $\bar{R}_{i\bullet}$ is the group's mean rank, and N is the total number of observations in the study. Given that in the present study $N = 12$, the estimated relative treatment effect of Group A would be $(5.375 - .5)/12 = .41$.

The interpretation of relative treatment effects is similar to the interpretation of means. If the groups being compared have similar relative treatment effects, this can

Table 1
Example Calculations of Relative Treatment Effects for the Analysis-of-Variance-Type Statistic

	Group A				Group B				Group C			
Original scores	11	12	17	18	10	11	13	15	20	21	23	25
Corresponding rank scores	2.5	4	7	8	1	2.5	5	6	9	10	11	12
Sum of ranks	21.50				14.50				42.00			
Mean rank	5.38				3.63				10.50			
Relative treatment effect	0.41				0.26				0.83			

be interpreted as indicating that the groups do not differ much (in terms of participants' typical response on the dependent variable). In contrast, large differences in relative treatment effects suggest that the groups differ significantly. Relative treatment effects for the current data set are shown in Table 1. Participants in Group C ($\hat{p}_C = .83$) tend to have higher scores on the dependent variable than do participants in Groups A ($\hat{p}_A = .41$) or B ($\hat{p}_B = .26$). The ATS can be used to determine whether these differences are statistically significant. In the present case, the null hypothesis that the groups have equivalent relative treatment effects is rejected, ATS(1.81, 7.42) = 11.38, $p = .006$. The values in parentheses (1.81 and 7.42) refer to the degrees of freedom for the test.

The ATS can analyze data from independent groups, repeated measures, and mixed designs. Brunner et al. (2002) developed macros that allow the easy generation of the ATS in SAS. Most of the macros are now also available for R. They can be downloaded from http://www.ams.med.uni-goettingen.de/de/sof/ld/makros.html. The Web page is written in German; however, the macros are clearly labeled and downloading them is straightforward. Brunner et al. (2002) provided instructions for using the SAS macros. Shah and Madden (2004) is another useful source of information, targeted at applied researchers unfamiliar with the ATS rather than statisticians. The article is accompanied by a comprehensive electronic supplement that illustrates how to use the SAS macros to generate and interpret the ATS and relative treatment effects. Wilcox (2003, 2005) also provided coverage of the ATS, which he referred to as the Brunner, Dette, and Munk (BDM) method. Wilcox discussed code for calculating the ATS using R. He also discussed how to follow up significant ATS main effects and interactions with multiple comparison procedures.

Other Rank-Based Methods

The ATS is just one of many modern robust rank-based methods. A prominent rank-based approach to ANOVA and regression is that of Hettmansperger and McKean (1998). Their approach is sometimes called *Wilcoxon analysis* (WA). WA evaluates hypotheses analogous to those assessed by classic parametric methods. For example, the null hypothesis tested in regression (i.e., no relationship between the predictors and the criterion variable; beta weights of zero) is exactly the same in WA as it is in ordinary least squares regression. The difference between the two procedures is that they use different methods to fit the regression line. In ordinary least squares regression, the regression line minimizes the sum of the squared residuals. A single outlier can substantially alter the slope of the regression line, reducing its fit to the data. In contrast, WA, which involves ranking residuals, minimizes the impact that extreme criterion (*y*-axis) scores have on the regression line. The result is that the WA line often provides a better fit to data than does the least squares line. It is important to note that although WA is robust to outliers in the *y*-space, it is not robust to extreme predictor (*x*-axis) values (neither is ordinary least squares regression). In such situations, a modified version of WA called *weighted Wilcoxon techniques* (WW) can be used. WW ensures that analyses are robust to outliers in both the *x*- and *y*-spaces. However, it is preferable to use WA in situations where there are no outliers in the *x*-space, as WA is more powerful than WW. See Hettmansperger and McKean (1998) and McKean (2004) for more details.

Free Web-based software known as RGLM can conduct WA via the Internet. RGLM is located at http://www.stat.wmich.edu/slab/RGLM. To conduct an analysis, a user uploads data to RGLM or enters data into a form. Analysis options are selected, and RGLM then conducts the analysis and outputs the results. The interface can perform WA simple and multiple regression, as well as WA alternatives to the single and paired samples *t* tests, one-way and factorial ANOVAs, and analysis of covariance. An appealing aspect of RGLM is that both WA and classic parametric analyses are reported side-by-side. This allows users to observe whether the two procedures produce equivalent or conflicting results. Abebe, Crimin, and McKean (2001) and Crimin, Abebe, and McKean (in press) provided a guide to conducting robust analyses using RGLM. It is important to note that the RGLM interface only conducts WA. Users wishing to conduct WW analyses should make use of the experimental site http://www.stat.wmich.edu/slab/HBR2. It is also possible to conduct WA and WW analyses using R. Terpstra and McKean (2005) provided instructions for carrying out WW analyses using R. R code for WA and WW is available for download from http://www.jstatsoft.org/v14/i07 and http://www.stat.wmich.edu/mckean/HMC/Rcode.

Readers interested in further information about rank-based methods may like to consult Higgins (2004). Journals that regularly feature articles about modern robust methods include *Psychological Methods*, *Educational and Psychological Measurement*, and the *Journal of Modern Applied Statistical Methods*. Articles in these journals are sometimes accompanied by useful software and instructions. For example, Serlin and Harwell (2004) published an article in *Psychological Methods* containing SPSS syntax that researchers can use to conduct a range of nonparametric approaches to regression.

Effect Size

The *Publication Manual of the American Psychological Association* (American Psychological Association, 2001) and many journals encourage researchers to report estimates of effect sizes in addition to statistical significance tests. Effect size provides information about the magnitude of an effect, which can be useful in determining whether it is of practical significance. Unfortunately, the most commonly reported effect sizes (e.g., Cohen's d, η^2) are predicated on the same restrictive assumptions (e.g., normality, homoscedasticity) as classic parametric statistical significance tests. Standard parametric effect sizes are not robust to violation of these assumptions (Algina, Keselman, & Penfield, 2005a; Grissom & Kim, 2001; Onwuegbuzie & Levin, 2003). Furthermore, using classic methods to calculate a confidence interval around the point estimate of an effect size with violated assumptions can lead to inadequate probability coverage (Algina, Keselman, & Penfield, 2005b, 2006b). In other words, a researcher may believe that he or she has formed a 95% confidence interval around the point estimate of an effect size, when in fact the degree of confidence may be lower (e.g., 85%). The take-home message is that researchers should not report estimates of standard effect sizes (nor confidence intervals around these estimates) if parametric test assumptions are violated, as the estimates and associated confidence intervals could be misleading. Fortunately, several robust alternatives to classic effect size indices exist.

A popular measure of effect size in the population is the standardized mean difference:

$$\delta = (\mu_A - \mu_B) / \sigma, \tag{3}$$

which is estimated by

$$(M_A - M_B)/SD, \tag{4}$$

where M_A is the mean of Group A and M_B is the mean of Group B. Variants of the standardized mean difference include Cohen's d, Glass's Δ, and Hedges's g (each variant uses a slightly different method to calculate the standard deviation). Robust analogues of the standardized mean difference exist to calculate the magnitude of an effect between two independent groups (Algina et al., 2005a; Algina, Keselman, & Penfield, 2006a) and two correlated groups (Algina et al., 2005b). Free software available from http://plaza.ufl.edu/algina/index.programs.html can compute these robust effect sizes. An attractive feature of the software is that it calculates accurate bootstrapped confidence intervals around the point estimate of the effect size (see also Keselman et al., 2008).

Another robust effect size is the probability of superiority (*PS*). *PS* has also been called the probabilistic index, intuitive and meaningful effect size index, area under the receiver operator characteristic curve, and the measure of stochastic superiority (Acion, Peterson, Temple, & Ardnt, 2006; Grissom, 1994; Grissom & Kim, 2005; Kraemer & Kupfer, 2006; Vargha & Delaney, 2000). *PS* is the probability that a randomly sampled score from one population is larger than a randomly sampled score from a second population. For example, imagine that a researcher wanted to compare men (Population 1) and women (Population 2) in terms of their height (dependent variable). If $PS = .70$, the probability that a randomly sampled man is taller than a randomly sampled woman is .70.

PS is easily estimated using software such as SPSS and SAS. First, run the Mann–Whitney U test and obtain the U value. In SPSS, the test is accessed by clicking *Analyze*, followed by *Nonparametric Tests* and *Two Independent Samples*. In SAS, use the NPAR1WAY procedure. Once the U value is obtained, estimate *PS* using the formula

$$PS_{est} = U/mn, \tag{5}$$

where U is the Mann–Whitney U statistic, m is the number of participants in the first sample, and n is the number of participants in the second sample (Acion et al., 2006; Grissom & Kim, 2005). For example, imagine $U = 80$, $m = 10$, and $n = 20$. Substituting these values into the formula above, we get

$$PS_{est} = 80/(10 \times 20) = .40. \tag{6}$$

The calculation of *PS* values using the above formula is only appropriate for independent groups designs. For a variant of *PS* that is appropriate for repeated measures designs, see Grissom and Kim (2005, pp. 114–115) or Vargha and Delaney (2000).

It is possible to compare *PS* values with those that would be obtained under normal theory using other estimates of effect size, such as Cohen's d. Grissom (1994) presented a comprehensive table of d values ranging from 0 to 3.99 and corresponding *PS* values. Using the table, it is possible to establish that $d = 0$ (i.e., no difference between group means) is equivalent to $PS = .50$. A small effect size ($d = .20$) is equal to a *PS* of .56, a medium effect size ($d = .50$) is equivalent to $PS = .64$, and a large effect size ($d = .80$) is equivalent to $PS = .71$. Grissom and Kim (2005, p. 109) provided a table for converting between d, *PS*, and the population point-biserial correlation, r_{pb}. *PS* can also be converted to the number needed to treat (NNT), an effect size index that is particularly appropriate for conveying information in psychotherapy outcome studies or other behavioral research that involves comparisons between treatments (or treatment and control or placebo conditions). NNT is defined as the number of patients that would need to be treated with Treatment A to experience one greater treatment success than if the same

number of patients were treated with Treatment B. For example, imagine a randomized controlled trial in which cognitive behavior therapy is compared with psychoeducation for the treatment of depression. Success is defined as the remission of depression at posttreatment. An NNT of 3 would indicate that it is necessary to treat three patients with cognitive behavior therapy, to have one more patient remit than if the same number of patients were treated with psychoeducation.

Many useful resources provide further information about the robust standardized mean difference, *PS*, NNT, and other effect size measures. Grissom and Kim (2005) is an authoritative source of information about numerous effect sizes for use in a wide range of designs (including factorials). Wilcox (2003, 2005) discussed various effect size measures and provided software for R used to form confidence intervals around *PS*. Kromrey and Coughlin (2007) prepared a SAS macro used to calculate *PS*, Algina's robust standardized mean difference, and a range of other robust effect sizes. Kraemer and Kupfer (2006) discussed the estimation of *PS* and NNT when using dichotomous rather than ordinal or continuous dependent variables. For situations (such as meta-analysis) in which original data are not available, the estimation of *PS* may not be possible. In these cases, *PS* is estimated using McGraw and Wong's (1992) common language effect size statistic.

Summary

Most researchers analyze data using outdated methods. Classic parametric tests, effect sizes, and confidence intervals around effect size statistics are not robust to violations of their assumptions, and violations seem to occur frequently when real data are analyzed. Researchers relying on statistical tests (e.g. Levene's test) to identify assumption violations may frequently fail to detect deviations from normality and homoscedasticity that are large enough to seriously affect the Type I error rate and power of classic parametric tests. We recommend that researchers bypass classic parametric statistics in favor of modern robust methods. Modern methods perform well in a much larger range of situations than do classic techniques. The use of modern methods will result in researchers finding more statistically significant results when real effects exist in the population. Using modern methods will also reduce the number of Type I errors made by researchers and result in more accurate confidence intervals around robust effect size statistics. A range of accessible texts about modern methods is available (e.g., Wilcox, 2001, 2003), as well as a wide range of software to perform modern analyses. Given the wealth of resources available, researchers have a tremendous opportunity to engage in modern robust statistical methods.

REFERENCES

Abebe, A., Crimin, K., & McKean, J. W. (2001). Rank-based procedures for linear models: Applications to pharmaceutical science data. *Drug Information Journal, 35,* 947–971.

Acion, L., Peterson, J. J., Temple, S., & Ardnt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine, 25,* 591–602.

Akritas, M. G., Arnold, S. F., & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association, 92,* 258–265.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005a). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10,* 317–328.

Algina, J., Keselman, H., & Penfield, R. (2005b). Effect sizes and their intervals: The two-level repeated measures case. *Educational and Psychological Measurement, 65,* 241–258.

Algina, J., Keselman, H. J., & Penfield, R. D. (2006a). Confidence interval coverage for Cohen's effect size statistic. *Educational and Psychological Measurement, 66,* 945–960.

Algina, J., Keselman, H., & Penfield, R. (2006b). Confidence intervals for an effect size when variances are not equal. *Journal of Modern Applied Statistical Methods, 5,* 2–13.

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

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57,* 49–64.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40,* 318–335.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician, 31,* 147–150.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144–152.

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

Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments.* New York: Wiley.

Brunner, E., & Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers, 42,* 1–52.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 121–129.

Crimin, K., Abebe, A., & McKean, J. W. (in press). Robust general linear models and graphics via a user interface. *Journal of Modern Applied Statistical Methods.* (Available from Ash Abebe at abebeas@auburn.edu or from Joe McKean at joseph.mckean@wmich.edu)

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., et al. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18,* 230–232.

D'Agostino, R. (1986). Tests for the normal distribution. In R. B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 367–420). New York: Dekker.

Fahoome, G., & Sawilowsky, S. S. (2000, April 24–28). *Review of twenty nonparametric statistics and their large sample approximations.* Paper presented at the annual meeting of the American Education Research Association. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED441031

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.

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

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79,* 314–316.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68,* 155–165.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135–146.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Erlbaum.

Harwell, M. R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics, 17,* 297–313.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17,* 315–339.

Hettmansperger, T. P., & McKean, J. W. (1998). *Robust nonparametric statistical methods.* London: Arnold.
Higgins, J. J. (2004). *Introduction to modern nonparametric statistics.* Pacific Grove, CA: Brooks/Cole.
Hultsch, D. F., MacDonald, S. W. S., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 57,* P101–P115.
Jaccard, J., & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical child and adolescent psychology: Advanced issues and recommendations. *Journal of Clinical Child Psychology, 31,* 278–294.
Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson.
Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13,* 110–129.
Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.
Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry, 59,* 990–996.
Kromrey, J. D., & Coughlin, K. B. (2007, November). *ROBUST_ES: A SAS macro for computing robust estimates of effect size.* Paper presented at the annual meeting of the SouthEast SAS Users Group, Hilton Head, SC. Retrieved from http://analytics.ncsu.edu/sesug/2007/PO19.pdf
Leech, N. L., & Onwuegbuzie, A. J. (2002, November). *A call for greater use of nonparametric statistics.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED471346
Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.
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.
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361–365.
McKean, J. W. (2004). Robust analysis of linear models. *Statistical Science, 19,* 562–570.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.
Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 539–543.
Onwuegbuzie, A. J., & Levin, J. R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead us? To no good effect. *Journal of Modern Applied Statistical Methods, 2,* 131–151.
Ramsey, P. H. (1980). Exact Type I error rates for robustness of Student's *t* test with unequal variances. *Journal of Educational Statistics, 5,* 337–349.
SAS Institute. (2004). *SAS/STAT 9.1 user's guide*. Retrieved February 1, 2008, from http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/stat_ug_7313.pdf
Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60,* 91–126.
Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111,* 352–360.
Serlin, R. C., & Harwell, M. R. (2004). More powerful tests of predictor subsets in regression analysis under nonnormality. *Psychological Methods, 9,* 492–509.
Shah, D. A., & Madden, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology, 94,* 33–43.
Taylor, D. H. (1965). Latency models for reaction time distributions. *Psychometrika, 30,* 157–163.
Terpstra, J. T., & McKean, J. W. (2005). Rank-based analyses of linear models using R. *Journal of Statistical Software, 14.* Retrieved from http://www.jstatsoft.org/v14/i07
Toothaker, L. E., & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational and Behavioral Statistics, 19,* 237–273.
Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the *CL* common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132.
Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.
Wilcox, R. R. (2001). *Fundamentals of modern statistical methods.* New York: Springer.
Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* San Diego, CA: Academic Press.
Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.
Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA *f, w* and *f* statistics. *Communications in Statistics: Simulation and Computation, 15,* 933–943.
Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8,* 254–274.
Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67,* 55–68.
Zimmerman, D. W. (2000). Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *Journal of General Psychology, 127,* 354–364.