Behavior Research Methods, Instruments, & Computers
1998, 30 (4), 707-712

Comparing means from nonnormal distributions: The bisquare-weighted analysis of variance

REBECCA ANNE REGETH and W. WREN STINE
University of New Hampshire, Durham, New Hampshire

A new procedure analogous to the analysis of variance (ANOVA), called the bisquare-weighted ANOVA (bANOVA), is described. When a traditional ANOVA is calculated, using samples from a distribution with heavy tails, the Type I error rates remain in check, but the Type II error rates increase, relative to those across samples from a normal distribution. The bANOVA is robust with respect to deviations from a normal distribution, maintaining high power with normal and heavy-tailed distributions alike. The more popular rank ANOVA (rANOVA) is also described briefly. However, the rANOVA is not as robust to large deviations from normality as is the bANOVA, and it generates high Type I error rates when applied to three-way designs.

Often, the analysis of data in the behavioral sciences involves the calculation of an analysis of variance (ANOVA), in which it is assumed that the data come from normal distributions (Kirk, 1995, chap. 3, p. 97). However, normal distributions are rare (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986, chap. 1; Micceri, 1989; Mosteller & Tukey, 1977, chaps. 1C & 1D; Rousseeuw & Leroy, 1987, chap. 2; Stigler, 1977; Walberg, Strykowski, Rovai, & Hung, 1984), with outliers commonly occurring. We will describe two techniques for dealing with nonnormal sampling distributions when calculating ANOVAs: the rank ANOVA (rANOVA; Conover & Iman, 1981) and the bisquare-weighted ANOVA (bANOVA; Regeth & Stine, 1993, 1996).

The ANOVA procedure is not robust with respect to deviations from normality (Hampel et al., 1986, pp. 31-33). Specifically, the result of conducting an ANOVA on a heavy-tailed distribution (increased incidence of outliers) is a large drop in power (i.e., a rise in Type II errors; Regeth & Stine, 1993), relative to other procedures.

The ANOVA is based on minimizing error variance. The variance is extremely sensitive to outliers, which makes it a poor choice as a measure of error variability with outlier-prone, heavy-tailed distributions. Deviations from normal distributions toward those with heavy tails are not usually problematic, if they are recognized. However, it is often difficult to recognize that a distribution is nonnormal (Hampel et al., 1986, chap. 1; Regeth & Stine, 1993).

TRADITIONAL METHODS OF DEALING WITH NONNORMAL DISTRIBUTIONS

If detected, there are a few ways that nonnormal distributions can be treated. Data from a nonnormal distribution can be transformed to approximate a normal distribution more closely. However, mathematical transformations, such as a logarithmic transformation, may lead to: (1) altered hypotheses (Tabachnick & Fidell, 1989, p. 83), (2) the introduction of outliers (Tabachnick & Fidell, 1989, p. 83), (3) obscured interactions (Anderson, 1961; Kirk, 1995, pp. 103–104), and (4) new interactions (Anderson, 1961; Kirk, 1995, pp. 103–104). For more information on the controversy surrounding the use of transformations, see Games (1983, 1984), Games and Lucas (1966), and Levine and Dunlap (1982, 1983).

Another option for dealing with nonnormal distributions is to hand-filter outliers. Extreme values may be replaced with new data (by replicating the study), corrected (if possible), or Winsorized (by replacing the highest and lowest extreme data points with the next highest or lowest data points; Kirk, 1995, p. 169). There are specific statistical rules that may help data analysts determine whether values are extreme (see, e.g., Grubbs, 1969). However, these techniques work poorly with the sample sizes one typically has available (Hampel et al., 1986, chap. 1; Regeth & Stine, 1993).

It is often difficult to determine whether extreme values are the result of a mistake in data computation or collection. A gross error can alter the statistical analysis considerably (Hampel et al., 1986, pp. 25–28). There is a tradeoff between leaving extreme values alone (and letting them possibly contaminate the results) and throwing out or replacing extreme values (that may be legitimate scores in the population.) The problem is that most heuristics regarding the deletion of outliers are somewhat arbitrary, and hand-filtering outliers tends to reduce efficiency (Hampel et al., 1986, p. 70). In fact, researchers tend to overtrim "outliers" from normal distributions (Hampel et al., 1986, chap. 1).

A third approach is to use a nonparametric test (Blair, 1981; Blair & Higgins, 1980; Boneau, 1962). The Mann-Whitney *U* test or the Kruskal-Wallis one-way test may be used in lieu of a *t* or ANOVA test for one-way designs.

Correspondence concerning this article should be addressed to R. A. Regeth, Department of Psychology, Box 13046, SFA Station, Stephen F. Austin State University, Nacogdoches, TX 75962-3046 (e-mail: rregeth@sfasu.edu).

Furthermore, for two-way designs (involving an interaction), the ANOVA test can be run, using the rank orderings of data points rather than the actual scores (Conover & Iman, 1981). Data from different groups are combined and rank ordered, and the ANOVA is then calculated on the ranks rather than on the actual scores.

Nonparametric designs, such as the rank ANOVA (rANOVA), are only slightly less powerful than their parametric counterparts when the samples are drawn from a normal distribution (Blair, 1981; Blair & Higgins, 1980; Boneau, 1962; Regeth & Stine, 1993). These techniques are also relatively robust when the samples are chosen from slightly heavy-tailed distributions (e.g., when the sampling distribution contains a 10% contamination at scale 10; Regeth & Stine, 1993). However, the rANOVA procedure's power is reduced relative to the bisquare-weighted ANOVA (bANOVA) when they are calculated on samples from distributions with moderately heavy tails (e.g., sampling distributions with 20% contamination at scale 10; Regeth & Stine, 1993), and the rANOVA is subject to a large number of Type I errors when used with three-way designs (Sawilowsky, Blair, & Higgins, 1989).

THE BISQUARE-WEIGHTED ANOVA

The bANOVA's (Regeth & Stine, 1993) power for one-way and two-way designs is comparable with that of nonparametric and ANOVA tests when samples come from normal distributions, but it is more robust than both tests with respect to severe deviations from normality toward heavy tails.

In our Monte Carlo study (Regeth & Stine, 1993), three distributions were used to assess the usefulness of the bANOVA technique: a normal distribution, a distribution that looks normal but with heavy tails (a high incidence of outliers), and a distribution with an inflated variance to match that with heavy tails.

The bANOVA and ANOVA statistical tests were computed on 1,000 randomly drawn samples from these distributions. Sample sizes used were 7, 15, or 30 elements per cell. Effect sizes were also varied, so that there was either a small (.10), medium (.25), or large (.40) effect size. In addition, both one-way (CR-4; Kirk, 1995, chap. 5) and 2 × 2 factorial (CRF-22; Kirk, 1995, chap. 9) designs were tested.

Figure 1 shows the power for a medium effect size at an α-level of .05 for a typical CR-4 design. The first column shows the power for a normal distribution. The second and third columns show the power for the heavy-tailed and variance-inflated distributions, respectively. Notice that the ANOVA has only slightly more power than the bANOVA when calculated on data from a normal distribution (the dark gray columns) and from a variance-inflated distribution (the light gray column). However, the bANOVA has greater power when calculated on a distribution with heavy tails (the white column).

The results of the Monte Carlo study show that the ANOVA is extremely sensitive to outliers. It loses power when calculated on heavy-tailed distributions. However, the bANOVA performs very well on heavy-tailed distributions. It retains high power and low Type II error rates. In addition, the bANOVA performs almost as well as the ANOVA on data from normal distributions. The remain-

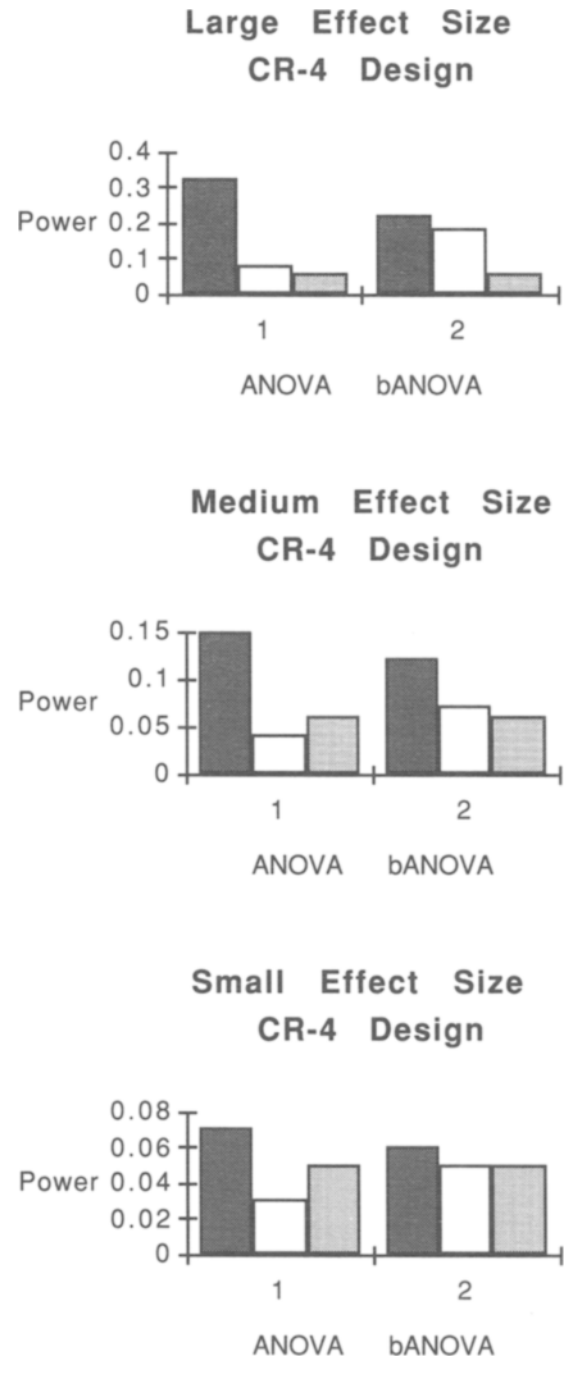

Figure 1. Results of Monte Carlo Study showing power of ANOVA and bANOVA. The dark gray, white, and light gray bars represent power for normal, heavy-tailed, and variance-inflated distribution, respectively.

der of this paper will illustrate the calculation of the bANOVA.

Formulas Used in Calculating the bANOVA

The bANOVA is calculated by first finding the bisquare-weighted average for each cell of the design. Hampel et al. (1986) present an overview of the bisquare-weighted average technique. The calculation is done through an iterative process. In the first iteration ($k = 1$), the initial estimate of the bisquare-weighted average is calculated, using the median as a measure of central tendency:

$$\underset{i=1,\ldots,n_j}{bw^{(0)}}\{X_{ij}\} = \underset{i=1,\ldots,n_j}{\text{Median}}\{X_{ij}\}. \quad (1)$$

The deviation of each score from the median is calculated.

Next, the median absolute deviation (MAD) is calculated:

$$\text{MAD}_j = \underset{i=1,\ldots,n_j}{\text{Median}}\left\{\left|X_{ij} - \underset{i=1,\ldots,n_j}{\text{Median}}\{X_{ij}\}\right|\right\}. \quad (2)$$

When the MAD is multiplied by 1.483, it provides a robust estimate of the standard deviation (Hampel et al., 1986, pp. 105 & 107). Next, the weights

$$\varepsilon_{ij}^{(k)} = \frac{X_{ij} - \underset{i=1,\ldots,n_j}{bw^{(k-1)}}\{X_{ij}\}}{1.483\text{MAD}_j} \quad (3)$$

and

$$w_{ij}^{(k)} = \begin{cases} \left(1 - \left(\dfrac{\varepsilon_{ij}^{(k)}}{r}\right)^2\right)^2, & \left|\varepsilon_{ij}^{(k)}\right| < r \\ 0, & \left|\varepsilon_{ij}^{(k)}\right| \geq r \end{cases} \quad (4)$$

are used to calculate the weighted average for $k = 1$:

$$\underset{i=1,\ldots,n_j}{bw^{(k)}}\{X_{ij}\} = \frac{\sum_{i=1}^{n_j} w_{ij}^{(k)} X_{ij}}{\sum_{i=1}^{n_j} w_{ij}^{(k)}}. \quad (5)$$

In the next iteration ($k = 2$), the bisquare-weighted average from the first iteration is used as the measure of central tendency rather than the median. New weights are found, using the same procedure as above. Then a new bisquare-weighted average is calculated from the new weights.

When the new bisquare-weighted average is approximately equal to the bisquare-weighted average from the previous iteration, the procedure is stopped. We use a difference of 0.001 between two successive bisquare-weighted averages as our criterion.

After the bisquare-weighted average and weights are found for each cell of the design, a weighted ANOVA can be calculated using SAS and the final weights (Equation 4) or some other statistical package. The last step involves transforming the F ratio from the weighted ANOVA into an F ratio for the bANOVA, using the following equation:

$$F_{bw} = (0.534 + 0.001206df_{\text{Error}})F. \quad (6)$$

The F_{bw} can then be compared to a critical F value found in most statistics books. The degrees of freedom from the weighted ANOVA are used to look up the critical F value.

The term *bisquare-weighted average* comes from the two squares in Equation 4. As is shown here, scores that are near the center of the distribution ($\varepsilon_{ij}^{(k)} \approx 0$) are given weights close to 1.0. Scores that are beyond approximately four robust estimates of the standard deviation ($\varepsilon_{ij}^{(k)} \geq r$) are given weights of zero. If a score is given a weight of zero, it indicates that the score is an outlier. However, scores that are given weights of zero during the first iteration may be given nonzero weights in subsequent iterations.

If sampling from a normal distribution, 99.994% of the scores will be within four standard deviations of the mean. Researchers may choose other values for r, depending on the relative consequences of Type I and Type II errors for their particular research program.

Example of the bANOVA Using Data

The calculation of a bANOVA will be described. We will use a one-way ANOVA design with three treatment groups and five subjects per group (see Table 1).

First, the median is found as a measure of central tendency (Equation 1). Deviations of the scores from the median are calculated. Then the median absolute deviation is determined (Equation 2). These deviations are then scaled, using Equation 3 (see Table 2).

After the scaled deviations are found, weights are calculated, using the weighting function in Equation 4. Re-

Table 1
One-Way ANOVA Design with Three Treatment Groups and Five Subjects per Group

	Group 1	Group 2	Group 3
	21	1	1
	22	2	2
	23	3	3
	24	4	4
	25	10	100
Median	23	3	3
Mean	23	4	22

Table 2
Scaled Median Absolute Deviations for Group 1

X_{i1}	$\varepsilon_{i1}^{(1)}$
21	−1.349
22	−0.674
23	0.000
24	0.674
25	1.349

Table 3
Weights for Group 1 for the First Iteration

X_{i1}	$\varepsilon_{i1}^{(1)}$	$w_{i1}^{(1)}$	$w_{i1}^{(1)}X_{i1}$
21	−1.349	0.786	16.506
22	−0.674	0.944	20.768
23	0.000	1.000	23.000
24	0.674	0.944	22.656
25	1.349	0.786	19.650
	$\sum_{i=1}^{5} =$	4.460	102.580

Table 4
First Iteration Calculations for Group 2

X_{i2}	$\varepsilon_{i2}^{(1)}$	$w_{i2}^{(1)}$	$w_{i2}^{(1)}X_{i2}$
1	−1.349	0.786	0.786
2	−0.674	0.944	1.888
3	0.000	1.000	3.000
4	0.674	0.944	3.776
10	4.720	0.000	0.000
	$\sum_{i=1}^{5} =$	3.674	9.450

call that this function allows us to give extreme scores weights of zero, removing them from the sample.

Next, the weighted average is computed (Equation 5). This average is compared to the median. If the two are approximately equal (within 0.001), the procedure is completed for this cell. If not, weights are calculated again, using the weighted average instead of the median.

For Group 1, the weighted average was 23. The median was also 23, so the iteration is complete for this cell (see Table 3):

$$\underset{i=1,\ldots,5}{bw^{(1)}}\{X_{i1}\} = \frac{102.559}{4.459} = 23 = \underset{i=1,\ldots,5}{bw^{(0)}}\{X_{i1}\}.$$

The first iteration calculations for Group 2 are presented in Table 4. Notice that the weight for the last score in the sample is zero. This score was too extreme to keep (an outlier) and was removed from the sample during this iteration. However, it may not be zero in subsequent iterations.

In the next iteration, the weighted average is used as a measure of central tendency, rather than the median. Deviations of the scores from the weighted average are calculated, and new weights are obtained. The second iteration for Group 2 is shown in Table 5.

The weighted average from the first iteration was

$$\underset{i=1,\ldots,5}{bw^{(1)}}\{X_{i2}\} = 2.572.$$

The weighted average from the second iteration was

$$\underset{i=1,\ldots,5}{bw^{(2)}}\{X_{i2}\} = 2.510.$$

If two weighted averages differ by less than 0.001, there is no need to continue to iterate. The difference between these two subsequent weighted averages was 0.062; therefore, the procedure will be repeated again, using the weighted average from the second iteration as a measure of central tendency.

After the fifth iteration, the difference between the two subsequent weighted averages was less than 0.001. The final results for Group 2 are shown in Table 6.

Notice that the weight for the last score ($X_{52} = 10$) is zero. This score was considered an outlier (it was beyond four robust estimates of the standard deviation). As will be discussed later, the removal of this score will require an adjustment to the degrees of freedom.

Group 3 took five iterations before the difference between the two subsequent weighted averages was less than 0.001. Table 7 shows the results for Group 3. The resulting bANOVA calculated from the weights is shown in Table 8. In contrast, the ANOVA calculated from the original scores is shown in Table 9.

As can be seen, the two extreme scores in Groups 2 and 3 had a large impact on the ANOVA. However, these scores were given weights of zero and, therefore, are not reflected in the F_{bw} and the bANOVA summary table.

Description of the SAS Code for the bANOVA

The SAS code for calculating a bANOVA is presented in Regeth and Stine (1996). It is also available for downloading (http://pubpages.unh.edu/~wws). The routine has four sections. The first calculates a median and MAD for each group in the design. As there is little novel in the first step, it will not be described further. A bisquare-weighted average is calculated for each group during the second step. The weights from the second step are used for the bANOVA in the third step. The final step transforms the F ratios from the third step and prints the results.

The second step uses a nonlinear regression program from SAS (the NLIN procedure) to calculate a bisquare-

Table 5
Second Iteration Calculations for Group 2

X_{i2}	$\varepsilon_{i2}^{(2)}$	$w_{i2}^{(2)}$	$w_{i2}^{(2)}X_{i2}$
1	−1.060	0.864	0.864
2	−0.386	0.981	1.963
3	0.288	0.990	2.970
4	0.963	0.888	3.550
10	5.009	0.000	0.000
	$\sum_{i=1}^{5} =$	3.723	9.340

Table 6
Final Results for Group 2

X_{i2}	$\varepsilon_{i2}^{(5)}$	$w_{i2}^{(5)}$	$w_{i2}^{(5)}X_{i2}$
1	−1.012	0.876	0.876
2	−0.337	0.986	1.972
3	0.337	0.986	2.958
4	1.011	0.876	3.505
10	5.057	0.000	0.000
	$\sum_{i=1}^{5} =$	3.724	9.311

Table 7
Final Calculations for Group 3

X_{i3}	$\varepsilon_{i3}^{(5)}$	$w_{i3}^{(5)}$	$w_{i3}^{(5)}X_{i3}$
1	−1.012	0.876	0.876
2	−0.337	0.986	1.972
3	0.337	0.986	2.958
4	1.011	0.876	3.505
100	65.745	0.000	0.000
	$\sum_{i=1}^{5} =$	3.724	9.311

Table 8
bANOVA

Source	SS	*df*	*MS*	*F*bw	*p* <
Between	1,017.59	2	508.80	142.55	.0001
Within	35.69	10	3.57		
Total	1,053.28	12			

Table 9
ANOVA

Source	SS	*df*	*MS*	*F*	*p*
Between	1,143.33	2	571.67	0.894	.434
Within	7,670.00	12	639.17		
Total	8,813.33	14			

weighted average for each group. The program allows one to fit an arbitrary function (its derivatives with respect to each parameter must exist) of several predictor variables and one criterion variable to an appropriate data set, using a least-squares criterion. It also allows the user to weight the contribution of each case in the data set arbitrarily. To calculate a bisquare-weighted average, one fits a one-parameter model to each group with the bisquare weighting function. That is, we fit

$$X_{ij} = \beta_j + \varepsilon_{ij},$$

where the contribution of X_{ij} is weighted, using Equation 4. So, if we define

$$\varepsilon_{ij} = X_{ij} - \beta_j$$

and then minimize

$$\sum_{i=1}^{n_j} \varepsilon_{ij}^2,$$

the resulting estimate of β_j, which we will denote β_j, equals

$$\bar{X}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}.$$

However, by weighting each X_{ij}, using Equation 4, we get

$$\hat{\beta}_j = \underset{i=1,\ldots,n_j}{bw^{(k)}} \{X_{ij}\},$$

where convergence is achieved in k iterations. Notice that any nonlinear regression program that minimizes the sum-of-squared error and in which the user can weight the contribution of the different cases could be used as just described in order to calculate a bisquare-weighted average.

The weights from the bisquare-weighted averages are used to calculate a weighted ANOVA. We use the general linear model (GLM) procedure with the weight statement for this calculation. An omnibus F ratio produced by the GLM procedure is then transformed, using Equation 6, and printed with the weights in the final step of our algorithm. The result of the transformation, F_{bw}, can be compared with a tabled value with the degrees of freedom from the weighted ANOVA.

SUMMARY

The bANOVA is a technique that can be used in place of the ANOVA with nonnormal distributions. In addition, it has nearly as much power as the ANOVA when used with normal distributions, making it ideal when normality cannot be determined.

The bANOVA is calculated iteratively. First, an estimate of central tendency is found. For the first iteration, the median is used, but subsequent iterations are based on estimates of the bisquare-weighted average. Deviations from the scores to the measure of central tendency are determined, and the median absolute deviation is calculated. Scores are then given weights that depend on how far they are from the median absolute deviation. Next, a new bisquare-weighted average is calculated by finding the ratio of the sum of the weights times the scores, divided by the sum of the weights. The next iteration is based on the bisquare-weighted average. This process is continued until the difference between the bisquare-weighted averages from the previous and the current iterations is less than 0.001. A weighted ANOVA can then be calculated from the weights, either by hand or by SAS (as is shown in Regeth & Stine, 1996).

REFERENCES

ANDERSON, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin*, **58**, 305-316.

BLAIR, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." *Review of Educational Research*, **51**, 499-507.

BLAIR, R. C., & HIGGINS, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* statistic under various non-normal distributions. *Journal of Educational Statistics*, **5**, 309-335.

BONEAU, C. A. (1962). A comparison of the power of the *U* and *t* tests. *Psychological Review*, **69**, 246-256.

CONOVER, W. J., & IMAN, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, **35**, 124-129.

GAMES, P. A. (1983). Curvilinear transformations of the dependent variable. *Psychological Bulletin*, **93**, 382-387.

GAMES, P. A. (1984). Data transformations, power, and skew: A rebuttal to Levine and Dunlap. *Psychological Bulletin*, **95**, 345-347.

GAMES, P. A., & LUCAS, P. A. (1966). Power and the analysis of variance of independent groups on nonnormal and normally transformed data. *Educational & Psychological Measurement*, **16**, 311-327.

GRUBBS, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1-21.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J., & STAHEL, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

KIRK, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

LEVINE, D. W., & DUNLAP, W. P. (1982). Power of the *F* test with skewed data: Should one transform or not? *Psychological Bulletin*, **92**, 272-280.

LEVINE, D. W., & DUNLAP, W. P. (1983). Data transformation, power, and skew: A rejoinder to Games. *Psychological Bulletin*, **93**, 596-599.

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

MOSTELLER, F., & TUKEY, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

REGETH, R. A., & STINE, W. W. (1993, February). *Robust ANOVAs with contaminated normal distributions*. Paper presented at the 1993 Annual Meeting of the American Association for the Advancement of Science, Boston.

REGETH, R. A., & STINE, W. W. (1996). The bisquare weighted analysis of variance: A technique for nonnormal distributions. In D. E. Ewing & W. E. Stinson (Eds.), *Proceedings of the ninth annual northeast SAS users group conference* (pp. 677-685). Boston: MA: Northeast SAS Users Group.

ROUSSEEUW, P. J., & LEROY, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

SAWILOWSKY, S. S., BLAIR, R. C., & HIGGINS, J. J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics*, **14**, 255-267.

STIGLER, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, **5**, 1055-1078.

TABACHNICK, B. G., & FIDELL, L. S. (1989). *Using multivariate statistics* (2nd ed.). Northridge, CA: Harper Collins.

WALBERG, H. J., STRYKOWSKI, B. F., ROVAI, E., & HUNG, S. S. (1984). Exceptional performance. *Review of Educational Research*, **54**, 87-112.

(Manuscript received September 18, 1996;
revision accepted for publication July 10, 1997.)