Original Article

Variance Heterogeneity in Published Psychological Research

A Review and a New Index

John Ruscio and Brendan Roche

Psychology Department, The College of New Jersey, Ewing, NJ, USA

Abstract. Parametric assumptions for statistical tests include normality and equal variances. Micceri (1989) found that data frequently violate the normality assumption; variances have received less attention. We recorded within-group variances of dependent variables for 455 studies published in leading psychology journals. Sample variances differed, often substantially, suggesting frequent violation of the assumption of equal population variances. Parallel analyses of equal-variance artificial data otherwise matched to the characteristics of the empirical data show that unequal sample variances in the empirical data exceed expectations from normal sampling error and can adversely affect Type I error rates of parametric statistical tests. Variance heterogeneity was unrelated to relative group sizes or total sample size and observed across subdisciplines of psychology in experimental and correlational research. These results underscore the value of examining variances and, when appropriate, using data-analytic methods robust to unequal variances. We provide a standardized index for examining and reporting variance heterogeneity.

Keywords: variance, variance ratio, heterogeneity of variance, assumptions, robustness

When comparing means across groups using conventional parametric statistical tests such as t and F, two of the required assumptions are that population distributions are normal and that population variances are equal. These statistical assumptions are made for various reasons, not the least of which is to allow researchers to compare the results of hand calculations for an empirical sample to a tabled critical value based on a hypothetical sampling distribution. Computer-intensive resampling techniques such as randomization tests (Edgington, 1980) or bootstrap methods (Efron & Tibshirani, 1993) now enable investigators to generate empirical sampling distributions based on the data at hand, rather than assuming normality and equal variances to generate hypothetical sampling distributions. In addition, a substantial literature has introduced and urged the use of modern data-analytic methods that improve robustness to violations of parametric assumptions (e.g., Wilcox, 2001, 2003; Wilcox & Keselman, 2003). For example, comparing trimmed means using Winsorized variances (Keselman, Othman, Wilcox, & Fradette, 2004) or using an approximate degrees of freedom test (Keselman, Algina, Lix, Wilcox, & Deering, 2008; Lix & Keselman, 1995) can be robust alternatives to comparing means using the standard t-test. Generating empirical sampling distributions and performing more robust analyses are not mutually exclusive options. These approaches can complement one another effectively.

Despite methodological advances that can improve inferential accuracy under certain conditions, for the most part popular statistical software packages (e.g., SPSS, SAS) continue to implement classic tests in conventional ways. An investigator would need to create or obtain specialized programs to take advantage of modern data-analytic techniques (e.g., many are available for SAS, S-Plus, R, or Matlab). Until widely used statistical software enables (via options) or encourages (via default settings) the use of such techniques, it seems unlikely that many investigators will do so. In the meantime, mainstream research will continue to rely on the assumptions of normality and equal variances. How often might these be violated in potentially problematic ways?

Micceri (1989) published a landmark study pertinent to the normality assumption. He obtained 440 large samples of achievement and psychometric measures, chosen because one might expect distributions to approximate normality better for these measures than for many others (e.g., laboratory measures such as reaction time or demographic measures such as household income). Micceri found that each sample failed at least one test of normality at the $\alpha = .01$ significance level. Asymmetry, excess tail weight, multimodality, and digit preferences constituted common features of these data, raising questions about the soundness of assuming normality when performing statistical tests.

Whereas Micceri (1989) provided an extensive evaluation of the normality assumption, the literature on the equal-variance assumption provides only glimpses. There are reasons to expect variance heterogeneity in some, perhaps many, research areas. Grissom (2000) explains that one might expect a relatively large variance ratio (VR, calculated as the largest within-group variance divided by the smallest within-group variance) when comparing treatment to control groups (see also Grissom & Kim, 2005). For example, an effective treatment might yield floor or ceiling

DOI: 10.1027/1614-2241/a000034

effects (depending on the direction of coding) on the outcome measure, and therefore less variance in the treatment than the control group. Alternatively, there might be individual differences in the size of a treatment effect, which can increase variance within that group only. As Grissom (2000, p. 156) puts it, "variance is more than just a nuisance parameter."

Variance heterogeneity has been studied in nonexperimental contexts, too. For example, Johnson, Carothers, and Deary (2008) review a long-standing debate about whether sex differences in the variability of intelligence might help to explain why there is no mean difference, yet males are overrepresented at both the top and bottom of the distribution. Differences in variance across experimental conditions or intact groups can be interesting and important in their own right. Because the same statistical tests, with the same parametric assumptions, typically are used in experimental and correlational research that compares group means, it seems worthwhile examining whether variance heterogeneity is more common in one or the other of these types of research.

A number of papers have commented on the extent of variance heterogeneity observed in a relatively small sample of studies in a particular journal or research area. Wilcox (1987) examined sample variances for 14 studies with one-way ANOVA designs that were published over the course of a few years in the *American Educational Research Journal*. Among these studies, three exhibited VR > 16. In a review that included 86 dependent variables in educational research articles, Keselman et al. (1998) found a mean VR of 2.0 (SD = 2.6), with extreme VRs ranging up to 23.8. Grissom (2000) reported that VRs for 10 studies in an issue of the *Journal of Consulting and Clinical Psychology* all exceeded 3.2. The modal VR was 4 and the largest values ranged as high as 281.8. Each of these studies suggests that there may be a nontrivial number of violations of the equal-variance assumption.

These studies suggest that assuming equal variances may be unwise, but they are limited to relatively few samples and research areas. Characteristics of data from clinical or educational research may or may not be representative of the broader research domain. In the present study, we examine a fairly large number of samples spanning a diverse array of subdisciplines of psychological science. In addition, we compare variance heterogeneity observed in samples to what would be expected due to normal sampling error when population variances are equal. At present, it is not clear whether a representative sample of VRs exceeds 1 by more than would be expected by sampling error alone.

In addition to documenting the extent of variance heterogeneity, we compare this to the levels that are considered problematic for the use of conventional parametric statistics, such as t and F. Some of the earliest research on the robustness of these tests to unequal variances suggested that there was little cause for concern. Boneau (1960) and Box (1954) reported that observed Type I error rates remained acceptably close to nominal levels under most or all of the conditions that they studied. The results of subsequent research have not always been as reassuring. For example, whereas Box assumed that VR would seldom exceed 3 and therefore did not study the effects of values greater than 3, Tomarken and Serlin (1986) studied VRs of 6 and 12 and found that this can either inflate or deflate Type I error rates relative to the nominal level as well as reduce statistical power. How often are VRs greater than 3 observed in empirical data? Is the extent of variance heterogeneity typically observed in published research sufficient to affect Type I error rates?

Research on robustness to unequal variances also has shown that the nature and extent of its influence depends on other factors (for reviews, see Grissom, 2000; Maxwell & Delaney, 2004; Wilcox, 2001, 2003). Unequal variances tend to be more problematic when the number of groups is large, when groups are of unequal size, or when the total sample size is small. Statistical power tends to be weakened and Type I error rates become conservative when larger variances are paired positively (directly) with larger group sizes, and power tends to be strengthened (with liberal Type I error rates) when variances are paired negatively (inversely) with group sizes. When combined with violations of the normality assumption, even fairly small violations of the equal-variance assumption can affect Type I error rates and statistical power substantially. The wisdom of investigators' routine reliance on conventional parametric tests for differences in group means would be called into question if variance heterogeneity turned out to be the norm. On the other hand, if variance homogeneity were the norm, or if variance heterogeneity tended to occur only under otherwise favorable conditions (e.g., with a small number of equally large groups), then the criticisms of standard statistical practices would lose much of their force and implementing alternatives might not be necessary.

To address these issues, we reviewed articles published in leading psychological journals spanning a wide range of subdisciplines. Data reported for between-group comparisons were coded such that we could calculate several indices of variance heterogeneity: The traditional VR, a coefficient of variance variation (CVV) proposed by Box (1954), and a new index. The CVV and our new index are calculated using the variances of all groups, rather than only those with the largest and smallest variances, and we demonstrate empirically that the new index in particular is less sensitive than the usual VR (or the CVV) to the number of groups. Because each of these indices is based on sample variances, we generated artificial comparison data matched to the number and sizes of groups in the empirical data to examine the expected variance heterogeneity due to normal sampling error. Using this as a baseline for comparison, we were able to determine whether variance heterogeneity in published research exceeds what would be expected by sampling error alone. We examined levels of variance heterogeneity across a number of factors, including experimental versus correlational research, subdisciplines of psychology, numbers of groups, relative group sizes, and total sample sizes. Finally, we performed analyses of artificial comparison data to assess the extent to which the observed levels of variance heterogeneity would affect Type I error rates.

Method

Data Source

Articles appearing in nine journals published by the American Psychological Association were retrieved using the PsycArticles database. To represent a wide variety of subfields within psychology, issues from the following journals were examined: *Behavioral Neuroscience*, *Developmental Psychology*, *Health Psychology*, *Journal of Abnormal Psychology*, *Journal of Applied Psychology*, *Journal of Educational Psychology*, *Journal of Experimental Psychology* (though we refer to this as a single entity, it is operated and published as five separate journals), *Journal of Consulting and Clinical Psychology*, *and Journal of Personality and Social Psychology*. Because there were too few studies reporting descriptive statistics for between-group comparisons in *Behavioral Neuroscience*, *Journal of Applied Psychology*, and the *Animal Behavior Processes* volumes of the *Journal of Experimental Psychology*, these were dropped. For each of the remaining seven journals, issues were selected at random from recent volumes and articles were examined until a total of 65 was obtained for each journal.

To qualify for inclusion, an article had to report the *M*, *SD* (or variance, or *SE* of the *M*), and *n* for each of at least two groups in a study. For papers presenting multiple studies, the first study that reported the necessary descriptive statistics was coded, with the exception that data were not drawn from research identified as a pilot study. For studies with multiple dependent measures, the first one mentioned in the text of the results was used; if multiple measures were included in a table, one was chosen at random. Meta-analyses were not considered because the unit of analysis differs from primary studies.

The number of volumes and issues per year varied across journals, as did the proportion of articles that qualified for inclusion in this study. As a result, the sample of 455 articles spans issues from late 2002 through late 2007.

Coding

Several variables were coded for each of the 455 articles. Identifying information included the journal title, year, volume number, issue number, and article title. The total sample size N and the number of groups k were recorded along with the *SD* and *n* for each group. For factorial designs that met the reporting requirements for inclusion in the study, k was recorded as the number of levels of the factor for which data were recorded (e.g., if *SD* and *n* were reported for all four levels of one factor in a 4 × 2 design, $k = 4$). If variances or *SE*s of the *M* were reported, these were converted to *SD* units (*SD* = sqrt[variance], *SD* = sqrt[N] × *SE* of the *M*). Finally, a study was coded as experimental if group membership was manipulated via random assignment, otherwise it was coded as correlational. This classification refers only to the variable used to form the groups whose variances were compared in the present study – other aspects of the study may or may not have involved experimental manipulations.

Data were checked for accuracy in two ways. First, various calculations were performed to check for coding errors. For example, group *n*s were summed to ensure that the total matched the recorded total sample size N for each article, the number of groups for which data were recorded was checked against the recorded value of k, and original sources were revisited to confirm data for studies with unusually large VRs (i.e., VR > 20). Next, coding reliability was evaluated. The second author coded all articles in the study, and the first author randomly selected 10% of articles for coding and did so blind to values recorded by the second author. No coding discrepancies were observed.

Measures of Variance Homogeneity

To quantify the extent to which variances were homogeneous, three measures were calculated for each study. First, VR was calculated as the ratio of the largest variance to the smallest variance. Second, the CVV developed by Box (1954) and used by Rogan, Keselman, and Breen (1977) was calculated as follows:

$$\mathrm{CVV} = \sqrt{\frac{\sum\left(df_k \times \left(s_k^2 - s_p^2\right)^2\right)}{N-k}} \Big/ s_p^2,$$

where $s_p^2 = \frac{\sum\left(df_k \times s_k^2\right)}{\sum df_k}$, N is the total sample size, k is the number of groups, and $df_k = n_k - 1$.

Due to normal sampling error, these measures tend to increase with the number of groups even when population variances are homogeneous. To facilitate comparisons across studies with different numbers of groups we created a third measure, referred to as the standardized variance heterogeneity (SVH). The calculation of the SVH was accomplished in three steps:

1. Convert the variance for each group i into an adjusted proportion of the sum of all groups' variance: $a_i = \frac{k s_i^2}{s_T^2}$, where s_i^2 is the sample variance for group i, $s_T^2 = \sum s_i^2$, and k is the number of groups. Because the sum of these adjusted proportions equals k, the final index is less sensitive to the number of groups than is the usual VR.
2. Calculate the standard deviation of the a_i values, using k rather than $k - 1$ in the denominator.
3. Divide the value from step 2 by the square root of $k - 1$, which represents the maximum value obtainable in step 2 for a given k.

The resulting SVH can range from 0, which represents perfectly equal sample variances, to 1, which represents maximally heterogeneous sample variances. Table 1 provides illustrative *SD*s for varying numbers of groups that

Table 1. Illustrative standard deviations for a range of standardized variance heterogeneity (SVH) values

	SVH							
SD	.00	.05	.10	.20	.30	.50	.90	1.00
1	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
2	10.00	9.51	9.05	8.16	7.34	5.77	2.29	0.00
3	10.00	9.18	8.43	7.13	6.02	4.23	1.39	0.00
4	10.00	8.92	7.97	6.40	5.17	3.37	0.99	0.00
5	10.00	8.69	7.59	5.85	4.56	2.81	0.78	0.00

Note. The first *k* entries in a given column represent the *SD*s for *k* groups that yield the SVH in the column heading. For example, the first three entries in the column labeled ".20" (10.00, 8.16, and 7.13) represent the *SD*s for three groups that would yield SVH = 0.20. The minimum SVH of .00 is always obtained when all *SD*s are equal, and the maximum SVH of 1.00 is always obtained when all *SD*s but one are 0.

yield specified values of SVH. The first *k* entries in a given column represent the *SD*s for *k* groups that yield the SVH in the column heading. For example, the first three entries in the column labeled ".20" (10.00, 8.16, and 7.13) represent the *SD*s for three groups that would yield SVH = 0.20. The minimum SVH of 0 is always obtained when all *SD*s are equal, and the maximum SVH of 1 is always obtained when all *SD*s but one are 0.

A final note on the calculation and reporting of variance heterogeneity measures involves the possibility of erroneous reporting. Though we checked our own coding and data entry carefully, we do not assume error-free reporting in the original articles. Rather than attempting to identify and remove errors by censoring samples with extremely heterogeneous variances, we chose to take the reported values at face value rather than knowingly introduce a bias. We see no reason to expect reporting errors to inflate variance heterogeneity (e.g., mistakenly reporting a group's *SD* as a much smaller or larger value) more often than they deflate it (e.g., mistakenly reporting the same *SD* for more than one group), and if anything we would expect errors involving highly discrepant *SD*s to be corrected more readily than errors involving highly similar *SD*s. Identifying and removing scores from the upper tails of the distributions of VR, CVV, or SVH values would introduce a bias tending to underestimate the extent and consequences of variance heterogeneity. In any case, none of our conclusions rest on findings for a small number of samples that may represent reporting errors.

Artificial Comparison Data

In any sample of data, the VR will exceed 1, CVV will exceed 0, and the SVH will exceed 0 due to normal sampling error. To simulate the amount of sample variance heterogeneity expected when population variances are equal, artificial comparison data sets were generated. Using these comparison data, variance heterogeneity statistics could be calculated in data tailored to the characteristics of the empirical samples under study. For each empirical sample, comparison data were generated by drawing independent random values from populations with equal variances. The number of groups and sample size of each group were matched in the comparison data, and a number of such replications were performed for each empirical data set; 100 replications were used when comparing observed values to those expected by chance, and 10,000 replications were used when estimating Type I error rates. Because research reports seldom provide full score distributions, scores for comparison data sets were drawn from normally distributed populations. This itself has been shown to be an unrealistic assumption (Micceri, 1989), but generating comparison data required the specification of a population distribution and normality seemed a more reasonable choice than any particular nonnormal distribution. Moreover, using normal population distributions allowed us to examine the influence of variance heterogeneity in isolation rather than confounding it with the violation of another parametric assumption. Artificial data were generated using programs written for the R computing environment (available on request).

Results

The left portion of Table 2 summarizes the distribution of VR values for all articles. The table also breaks the results down by the number of groups, combining the relatively few studies with $k > 4$ into a single category. All distributions were positively skewed. Summary statistics include the minimum and maximum values, several percentile points in between, the *M*, and the percent of samples yielding VR > 3. Whereas in his pioneering work on variance heterogeneity, Box (1954) assumed that VRs would seldom exceed 3, the present results show that VRs exceeded 3 nearly one-quarter of the time (23.18%), with much larger values also fairly common. In contrast, among comparison data sets in which the number of groups and size of each group were matched to the values in the empirical samples, but scores were drawn from populations with homogeneous variances, sampling error alone yielded VR > 3 only 6.20% of the time. Across all 455 empirical samples, VR spanned values from 1.00 to 20,264.36; VR was undefined for two samples in which the smallest variance was 0. Despite the enormous range, the *Mdn* value was 1.64 and the middle 50% of values ranged from 1.23 to 2.76. VR values

Table 2. Summary of variance heterogeneity values

	Variance ratio					Coefficient of variance variation					Standardized variance heterogeneity				
	$k = 2$	$k = 3$	$k = 4$	$k \geq 4$	All	$k = 2$	$k = 3$	$k = 4$	$k \geq 4$	All	$k = 2$	$k = 3$	$k = 4$	$k \geq 4$	All
N	234	90[a]	93	36[a,b]	453	234	91	93	37[b]	455	234	91	93	37[b]	455
Min	1.00	1.09	1.08	1.13	1.00	0.00	0.04	0.02	0.04	0.00	0.00	0.03	0.02	0.02	0.00
10th percentile	1.02	1.18	1.37	1.35	1.08	0.01	0.07	0.10	0.12	0.04	0.01	0.05	0.07	0.05	0.03
25th percentile	1.12	1.36	1.53	1.82	1.12	0.05	0.13	0.17	0.20	0.10	0.06	0.10	0.10	0.11	0.08
50th percentile	1.33	1.92	2.03	2.74	1.64	0.14	0.27	0.28	0.33	0.20	0.14	0.19	0.16	0.16	0.16
75th percentile	1.84	3.11	3.15	5.10	2.76	0.28	0.45	0.42	0.44	0.40	0.30	0.32	0.25	0.20	0.29
90th percentile	3.72	5.90	5.75	9.43	5.10	0.56	0.57	0.54	0.64	0.57	0.58	0.43	0.34	0.33	0.51
Max	101.54	85.09	406.93	> 20K	> 20K	0.98	1.29	1.56	1.40	1.56	0.98	0.92	0.90	0.53	0.98
M	2.51	3.95	8.84	597.66	51.39	0.21	0.32	0.35	0.39	0.28	0.22	0.24	0.20	0.18	0.22
Percent > 3.00	15.38	27.78	30.11	44.44	23.18										
% positive pairings											31.62	37.36	30.11	45.95	33.63
% negative pairings											29.91	42.86	40.86	37.84	35.38
% unpaired											38.46	19.78	29.03	16.22	30.99

Note. k = number of groups; N = number of studies; "Percent > 3.00" is the percentage of studies with variance ratio (VR) > 3.00; "% positive pairings" is the percentage of studies for which there was a positive correlation between group sizes and variances; "% negative pairings" is the percentage of studies for which there was a negative correlation between group sizes and variances; "% unpaired" is the percentage of studies for which group sizes and/or variances were equal. Each distribution of VRs was positively skewed.

[a]The VR was undefined for one sample with $k = 3$ and one sample with $k = 6$ because the smallest variance was 0.

[b]There were 9 samples with $k = 5$, 19 samples with $k = 6$, 2 samples with $k = 7$, 6 samples with $k = 8$, and 1 sample with $k = 10$.

increased substantially with the number of groups, with a *Mdn* of 1.33 for $k = 2$ and a *Mdn* of 2.74 for $k > 4$. Across samples with $k = 2$, $k = 3$, $k = 4$, and $k > 4$, there was a statistically significant difference in VR values, $F(3, 449) = 4.38$, $p = .001$. Because the assumption of variance homogeneity was violated for these data, p was calculated using a randomization test (Edgington, 1980) with 10,000 random permutations; the same was done for each F-test that follows. Naturally, some portion of the increase in VR with larger k can be attributed to normal sampling error: The more groups, the more the extreme sample variances are likely to differ even if population variances are homogeneous.

The middle portion of Table 2 summarizes the distributions of CVV values. Across all samples, CVV values ranged from 0.00 to 1.56. The distributions of CVV values were skewed, but less so than for the VR. Like VR values, CVV values increased substantially with the number of groups, with a *Mdn* of 0.14 for $k = 2$ and a *Mdn* of 0.44 for $k > 4$. Across samples with $k = 2$, $k = 3$, $k = 4$, and $k > 4$, there was a statistically significant difference in CVV values, $F(3, 451) = 11.43$, $p < .001$.

The right portion of Table 2 summarizes the distributions of SVH values. Across all samples, SVH values extended from the smallest possible value of .00 nearly all the way to the largest possible value of 1.00; the observed maximum was .98. The distributions were positively skewed, but mildly. The *Mdn* value was 0.16, and the middle 50% of values ranged from .08 to .29. As intended, this standardized index was insensitive to differing numbers of groups. There was no statistically significant difference in the mean SVH across samples with $k = 2$, $k = 3$, $k = 4$, and $k > 4$, $F(3, 451) = 0.86$, $p = .413$.

Because the SVH is calculated using sample variances, due to normal sampling error it will deviate .00 even with population homogeneity. The analysis of artificial comparison data provides an interpretive benchmark that takes into account the number and sizes of groups for each empirical sample. Figure 1 shows the distributions of SVH values for the empirical data (solid curve) and the comparison data (dotted curve), along with summary statistics for each distribution. There is a substantial difference between these distributions, with larger SVHs more common among empirical samples than comparison data. To quantify the magnitude of this effect, the nonparametric effect size A was calculated (Ruscio, 2008). This is equivalent to the area under a receiver operating characteristic curve, or the probability that a value drawn at random from the distribution of empirical data SVHs is greater than a value drawn at random from the distribution of comparison data SVHs. Across all 455 samples, $A = .64$, and a 95% confidence interval (CI) constructed using the bias-corrected and accelerated bootstrap method (Efron & Tibshirani, 1993) ranged from .61 to .67. To help interpret these results, one can convert Cohen's d values of .20, .50, and .80 – which correspond to small, medium, and large effects according to Cohen's (1992) rule of thumb – to A values of .56, .64, and .71 (also note that $A = .50$ corresponds to $d = .00$). Thus, the SVHs for empirical data differ from those for comparison data to a statistically significant extent with an effect size considered medium by conventional standards. This suggests that in many of the studies reviewed here the homogeneity of variance assumption may not be satisfied, with violations of nontrivial magnitude.

Figure 2 presents the distribution of SVHs for empirical samples at each level of k, again accompanied by the

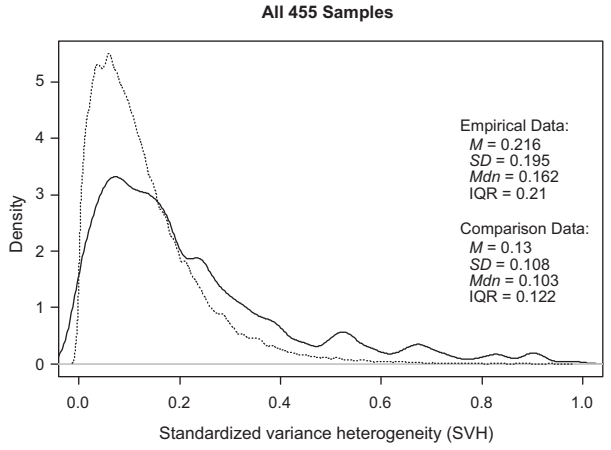

Figure 1. Standardized variance heterogeneity (SVH) values for all samples. The solid curve shows the results for empirical samples, and the dotted curve shows the results for artificial comparison data, with 100 replications per empirical sample. IQR, interquartile range.

distribution for comparison data and summary statistics. In each case, larger values are more common for empirical samples than for comparison data. The magnitude of the effect was closest to medium or large at each level of k, with A ranging from .60 to .70 across these four analyses; 95% CIs did not include the null value of .50. Table 3 extends these results across journals and type of study, classified as experimental versus correlational. SVHs were larger for the 283 experimental studies (Mdn = 0.17) than the 172 correlational studies (Mdn = 0.15), $t(453) = 2.43$, $p = .015$, though the effect size was relatively small, $A = .56$, and may be due to the fact that experimental studies had smaller group sizes (median $n = 28$) than correlational studies (median $n = 53$). Despite this difference in SVHs across type of study, SVHs tended to be larger for the empirical samples than for comparison data regardless of the type of study or the journal in which it was published. Some of the within-journal comparisons did not reach statistical significance owing to the relatively low statistical power of those analyses. Nonetheless, the homogeneity of variance assumption appears to be comparably tenuous in experimental and correlational studies published in a wide range of journals (Figure 3).

As noted earlier, it is well known that parametric statistical tests are more robust to unequal variances when there are fewer groups, groups are of equal size, and the total sample size is large. If unequal variances tended to occur under these conditions, the present findings would provide less cause for concern. Our database allowed us to examine these possibilities. Results bearing on the first issue have already been presented: There was no statistically significant difference in SVH values across the number of groups.

Is there a relationship between variance heterogeneity and relative group sizes? This can be tested most straightforwardly among the 234 samples with $k = 2$ groups by using the proportion of cases in the larger group as an index of group size heterogeneity. This ranges from .50 for equal group sizes to values approaching 1 for increasingly unequal group sizes. There was no statistically significant rank-order correlation between the proportion of cases in the larger group and the SVH, $r_S(232) = -.03$, $p = .597$. As shown in Figure 4 (top graph), even though the distributions of these variables were not normal, a scatterplot smoother did not suggest a nontrivial relationship between them. To include all 455 samples in a similar analysis, we constructed a measure of group size heterogeneity as $1 - n_H/n_M$, where n_H is the harmonic mean group size and n_M is the arithmetic mean group size. This index ranges from 0 to 1, with 0 representing equal group sizes and larger values representing unequal group sizes. This index, too, was uncorrelated with the SVH, $r_S(453) = -.03$, $p = .478$, and the bottom graph in Figure 4 shows that a scatterplot smoother did not suggest a nontrivial relationship. It appears that the equality of group sizes, by itself, is unrelated to equality of variances.

Next, is there a relationship between variance heterogeneity and total sample sizes? The answer here also appears to be "no," but the results leading to this conclusion are a bit more complex. As shown in Figure 5, there is a statistically significant rank-order correlation such that as sample size increases, SVH decreases, $r_S(453) = -.31$, $p < .001$. However, purely on the basis of sampling error one would expect sample variance heterogeneity to decrease with sample size. Therefore, the important question is whether the observed inverse relationship is stronger than what one would expect due to sampling error alone. In other words, controlling for the predictable influence of normal sampling error, did studies with larger samples tend to be the ones with more homogeneous variances? To address this question, we turned once again to results for artificial comparison data. The solid curve in Figure 5 shows the monotonic decreasing relationship between sample size and SVH for the empirical samples, generated using a locally weighted scatterplot smoother. When the same technique was applied to comparison data, the dashed curve in Figure 5 emerged. These curves parallel one another closely throughout their ranges, suggesting that the downward slope over sample size can be attributed to normal sampling error. The difference in elevation (or intercept) between these curves reflects the fact that SVHs were higher for empirical data than comparison data, which has been established already. As with the results for analyses based on the number of groups and relative group sizes, the results for total sample sizes provide no evidence that variance heterogeneity tends to occur under circumstances when its influence would be less problematic for parametric statistical tests.

Finally, and perhaps most important, does variance heterogeneity have consequences for parametric statistical tests? To examine this question, 10,000 replication samples of artificial comparison data were generated for each empirical sample and submitted to F tests. For each empirical sample, the observed Type I error rate was calculated as the proportion of the p values for the 10,000 analyses that fell below a nominal Type I error rate of $\alpha = .05$. To determine whether an observed Type I error rate differed from the nominal rate, 95% control limits were calculated. Across all 455 empirical samples, observed Type I error rates ranged from .001 to .259 (M = 0.055, Mdn = 0.052). A total of

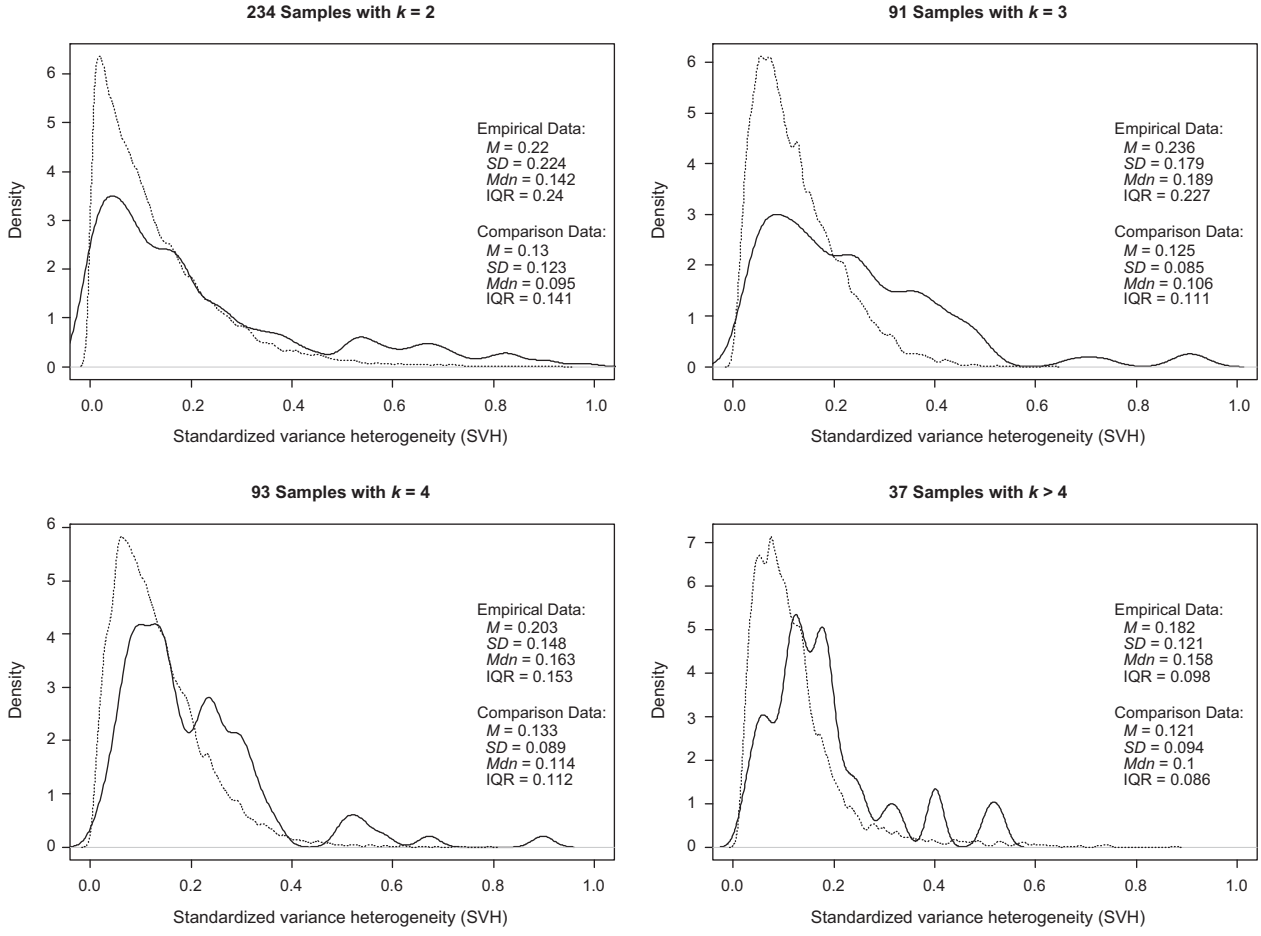

Figure 2. Standardized variance heterogeneity (SVH) values for samples with k = 2, 3, 4, or more than four groups. Each solid curve shows the results for empirical samples, and each dotted curve shows the results for artificial comparison data, with 100 replications per empirical sample. IQR, interquartile range.

Table 3. Median standardized variance heterogeneity (SVH) by journal and type of research

Journal	Experimental	*n*	Correlational	*n*	Total
Developmental Psychology	.16 (.11)	32	.14 (.08)[a]	33	.14 (.09)[a]
Health Psychology	.13 (.08)	34	.11 (.06)[a]	31	.13 (.07)[a]
Journal of Abnormal Psychology	.22 (.11)[a]	41	.21 (.09)[a]	24	.22 (.10)[a]
Journal of Consulting and Clinical Psychology	.12 (.09)[a]	47	.12 (.07)	18	.12 (.09)[a]
Journal of Educational Psychology	.16 (.12)[a]	36	.16 (.08)[a]	29	.16 (.10)[a]
Journal of Experimental Psychology	.22 (.17)	55	.17 (.17)	10	.19 (.17)
Journal of Personality and Social Psychology	.20 (.13)[a]	37	.18 (.10)[a]	28	.19 (.12)[a]
All journals	.17 (.12)[a]	283	.15 (.08)[a]	172	.16 (.10)[a]

Note. Each *Mdn* SVH for empirical samples is followed in parentheses by the *Mdn* SVH for artificial comparison data, with 100 replications per empirical sample.
[a]95% CI for the difference between SVHs for empirical samples and artificial data does not include the null value of A = .50.

45% of these values fell outside of the 95% control limits, 31% on the liberal side plus 14% on the conservative side. Generally, the conservative Type I error rates occurred when there were positive pairings between group sizes and variances and the liberal Type I error rates occurred when there were negative pairings. For the 314 samples with unequal group sizes and unequal variances, the sign of their correlation across groups was placed on the SVH to incorporate the direction of group size and variance pairings. Approximately equal numbers of samples had positive pairings (n = 153)

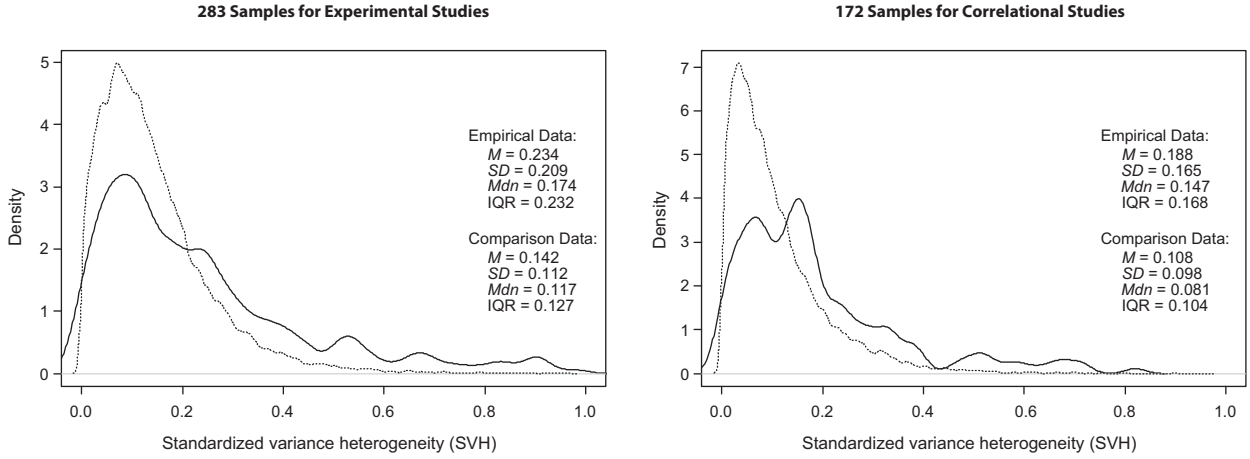

Figure 3. Standardized variance heterogeneity (SVH) values for studies classified as experimental versus correlational. Each solid curve shows the results for empirical samples, and each dotted curve shows the results for artificial comparison data, with 100 replications per empirical sample. IQR, interquartile range.

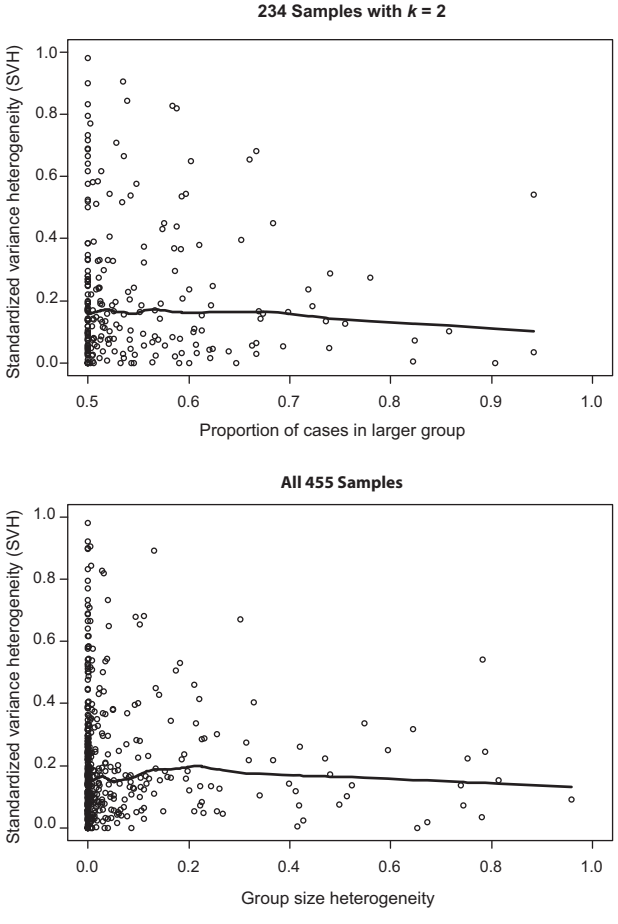

Figure 4. Standardized variance heterogeneity (SVH) by proportion of cases in the larger of two groups (top graph) or the group size heterogeneity (bottom graph). Solid curves were plotted using a locally weighted scatterplot smoother (LOWESS).

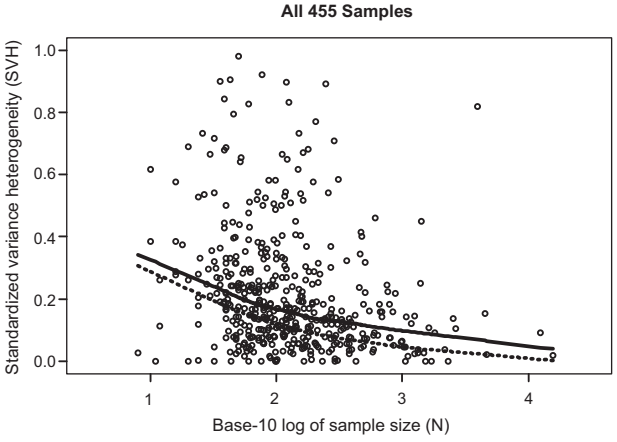

Figure 5. Standardized variance heterogeneity (SVH) by the base-10 logarithm of sample size (N); a log scale was used for sample size to reduce its extreme positive skew and obtain a more readable graph. The solid curve was plotted using a locally weighted scatterplot smoother (LOWESS). The dashed curve shows the LOWESS results for artificial comparison data, with 100 replications per empirical sample.

and negative pairings (n = 161); further details on pairings are summarized in Table 2. The rank-order correlation between this signed variant of the SVH and the observed Type I error rate was $r_S(312) = -.71$, $p < .001$. Thus, even though assumptions of independence and population normality were satisfied in these simulations, the variance heterogeneity observed in these empirical samples was sufficient to yield discrepancies between nominal and observed Type I error rates beyond what one would expect by chance. Whether the discrepancies were of a conservative or liberal nature was somewhat predictable when there was

either a positive or negative pairing between group sizes and variances.

Discussion

In general, sample variance heterogeneity appears to exceed what can be attributed to normal sampling error, and this can affect the Type I error rates of conventional parametric statistical tests for differences between group means. When comparing variance heterogeneity for all empirical samples and matched comparison data, the effect size of $A = .64$ means that there was a 64% chance that the SVH for a randomly selected empirical sample would exceed the SVH for a randomly selected comparison data set. This is elevated above the 50% chance that one would expect when population variances are homogeneous, suggesting that this assumption is violated in a nontrivial number of published studies. Comparable levels of variance heterogeneity were observed for studies with more or fewer groups, with equal and unequal group sizes, and with larger or smaller sample sizes. Experimental studies exhibited greater variance heterogeneity than correlational studies, though all or most of this apparent difference can be attributed to the smaller sample sizes in experimental research. Variance heterogeneity exceeded what would be expected by sampling error to a comparable extent in research published in the leading journals for developmental, health, abnormal, consulting/clinical, educational, experimental, and social/personality psychology.

Even when independent observations were drawn at random from normally distributed populations, analyses of artificial comparison data suggest that the variance heterogeneity observed in these 455 samples would be sufficient to affect Type I error rates quite often. The resulting bias can be either conservative or liberal, depending in part on whether group sizes and variances are paired positively or negatively. Variance heterogeneity does not appear to be more common under circumstances when its influence would be less problematic (e.g., a small number of equally large groups) than when it would be more problematic (e.g., a larger number of unevenly sized groups with a small total *N*). Our findings may provide a conservative estimate of the extent to which variance heterogeneity affects Type I error rates because our comparison data were drawn exclusively from normally distributed populations and allowed to vary along true ratio scales, whereas empirical data seldom approximate these ideals closely.

Data for this study were drawn from articles published in leading psychology journals, so it remains unknown whether comparable levels of variance heterogeneity would be observed in research published in less selective journals or appearing in other sources (e.g., books, conferences, unpublished manuscripts). We are not aware of any evidence that variance heterogeneity is detected or criticized in the peer-review process at top-tier (or other) journals. While acknowledging that this remains an empirical question, we predict that satisfaction of the equal-variance assumption is unlikely to differ much at lower-tier journals, in other sources, or in older or more recent research.

Like Micceri's (1989) work pertaining to normality, the present study found that data routinely violate a key assumption of conventional parametric statistical tests. What can researchers do about the possibility that variance heterogeneity will affect the Type I error rate or statistical power of tests between group means? At a minimum, we encourage investigators to examine the variances in their samples and either report them or provide a statistical summary. Whereas the usual VR is simple to calculate and easy to understand, Box's (1954) CVV fares more poorly on both counts and cannot be recommended for routine reporting. Though the SVH that we introduced is cumbersome to calculate, it does offer a few advantages relative to the VR. First, whereas there is no theoretical upper limit to the VR, the SVH is standardized in the sense that it is bounded by 0 (equal variances) and 1 (maximally heterogeneous variances). Second, when the smallest within-group variance is 0, the VR is undefined but the SVH can be calculated. For example, with sample *SD*s (or variances) of 0 and 1, the VR = 1/0, which is undefined, whereas the SVH = 1.00. With sample *SD*s of 0, 1, and 1, the VR is once again undefined and the SVH = 0.50. This hints at the third advantage, that for studies with more than two groups the SVH is a sufficient statistic whereas the VR is not. Because the VR is based on just two groups' variances (the largest and the smallest), it is insensitive to any other groups' variances; the SVH is calculated using each group's variance. Table 4 shows that the SVH can reflect important differences among samples that yield identical VR values. For example, whereas VR = 3 for samples whose group variances are {1, 1, 1, 3} or {1, 3, 3, 3}, their SVH values are .33 and .20, respectively. Finally, as demonstrated empirically, the SVH is less sensitive than the VR to the number of groups in a study. For each of these reasons, we suggest that the SVH may be a useful index of variance heterogeneity. At the same time, we developed this index in the context of a relatively simple research design, independent groups assessed using a single measure at a single time point. Whether this index or a variant on this theme will prove useful for more complex designs requires further study.

Of course, quantifying and reporting the extent of variance heterogeneity does not solve the potential problems that it can cause. Researchers can perform statistical tests of the null hypothesis of equal population variances to determine whether the conventional data-analytic method for comparing group means is a reasonable choice. For example, one can perform O'Brien's (1981) or Levene's (1960) test. Unfortunately, these and other available tests often have poor statistical power. Even with normal distributions, testing the equal-variance assumption can yield Type II errors, which provides unwarranted reassurance (Maxwell & Delaney, 2004; Wilcox, Charlin, & Thompson, 1986). Wilcox and Keselman (2003, p. 262) conclude that "tests for equal variances rarely have enough power to detect differences in variances of a magnitude that can adversely affect conventional methods." Even increasing α to unusually high levels for these tests yields poor statistical power. Keselman, Wilcox, Algina, Othman, and Fradette (2008) compared more than 600 procedures to test for variance heterogeneity between groups and recommend methods based

Table 4. Standardized variance heterogeneity (SVH) for six different samples with VR = 3

	A	B	C	D	E	F
	1	1	1	1	1	1
	3	1	3	1	1	3
	–	3	3	1	3	3
	–	–	–	3	3	3
SVH	0.50	0.40	0.29	0.33	0.29	0.20
p	.003	.000	.015	.000	.001	.045

Note. The entries in each column (sample) are the variances for each group in that sample. For each sample, 10,000 sets of artificial comparison data in which $n = 33$ scores per group (the median value observed in the 455 articles reviewed in the present study) were drawn from the standard normal distribution were generated and the SVH was computed for each. The *p* value was calculated as the proportion of the SVHs for comparison data that were greater than the SVH shown in the table. In each instance, variance heterogeneity was statistically significant at $\alpha = .05$. As usual, the statistical power of such a test depends on group sizes: *p* values would be lower for larger group sizes and higher for smaller group sizes.

on asymmetric trimming strategies – trimming different amounts of data from the upper and lower tails of distributions based on measures of skewness or tail length – to control Type I error rates; statistical power was not addressed in this study. As described in the note to Table 4, one can generate artificial comparison data to provide a sampling distribution under the null hypothesis of equal variances to calculate a *p* value for an index of variance heterogeneity such as the SVH. We doubt that this would provide a more powerful test than existing alternatives because the fundamental challenge involves the large sampling error for sample variances, but that might be worthy of empirical examination. In any case, calculating a *p* value for the SVH rather than performing another test has the virtue of being simpler to report. Rather than introducing an additional test statistic, one can append a *p* value to the existing presentation of an SVH value. A program written in R to calculate the SVH and its associated *p* value given only the size and variance (or *SD*) of each group in a study is available from the first author.

For any number of reasons, investigators might prefer not to rely on a conventional parametric statistical test to compare group means. Perhaps a test of the null hypothesis of equal variances yielded a statistically significant result. Even in the absence of such a result, in recognition of the poor statistical power of such a test and the fact that variance heterogeneity is common, researchers might choose to use alternative data-analytic techniques to compare scores across groups. The options include checking for and removing outliers (Wilcox & Keselman, 2003), performing nonlinear data transformations (Budescu & Appelbaum, 1981; Games, 1983; but note that this can affect the interpretation of interaction terms in a factorial model), using more robust measures of central tendency and variability (Wilcox & Keselman, 2003), using adjusted degrees of freedom tests (Keselman, Algina, et al., 2008; Lix & Keselman, 1995), and generating empirical sampling distributions rather than relying on hypothetical sampling distributions that necessitate more stringent parametric assumptions (Edgington, 1980; Efron & Tibshirani, 1993). Each of these approaches has its own strengths and weaknesses, and some can be used in complementary fashion. Though we did not specifically code for this when reviewing published research, we cannot recall any instances in which authors commented on variance heterogeneity or used statistical methods designed to handle it. The number of forgotten instances, if any, would still be very small.

Though we believe the present findings underscore the importance of considering one or more of the alternative approaches to conventional statistical tests, we do not believe that these findings have implications for how to make the wisest decisions. Having found that variance heterogeneity appears to be routine in psychological research, capable of affecting Type I error rates even if other assumptions are satisfied, we recommend that investigators learn about and implement data-analytic techniques that do not require – or are robust to violations of – the assumption of equal variances. Because the publications that we reviewed did not report raw data and it would not be feasible to obtain and perform secondary analyses of these data, we cannot empirically compare the available data-analytic alternatives to offer more specific guidance. Instead, we refer interested readers to the references cited above as starting points for exploring these approaches. The paper by Keselman, Algina, et al. (2008) provides an especially comprehensive treatment of a unified, robust approach, and an online supplement to that article contains program code for implementing it. Texts that provide excellent overviews of the pertinent issues include Grissom and Kim (2005), Maxwell and Delaney (2004), and Wilcox (2001, 2003).

References

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

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way model. *Annals of Mathematical Statistics, 25*, 290–302.

Budescu, D. V., & Appelbaum, M. I. (1981). Variance stabilizing transformations and the power of the *F* test. *Journal of Educational Statistics, 6*, 55–74.

Edgington, E. S. (1980). *Randomization tests*. New York, NY: Marcel Dekker.

Efron, B., & Tibshirani, R. (1993). *An introduction to bootstrap*. London, UK: Chapman and Hall.
Games, P. A. (1983). Curvilinear transformations of the dependent variable. *Psychological Bulletin, 93*, 382–387.
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. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.
Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science, 3*, 518–531.
Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (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., ... 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.
Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample *t* test. *Psychological Science, 15*, 47–51.
Keselman, H. J., Wilcox, R. R., Algina, J., Othman, A. R., & Fradette, K. (2008). A comparative study of robust tests for spread: Asymmetric trimming strategies. *British Journal of Mathematical and Statistical Psychology, 61*, 235–253.
Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin, 117*, 547–560.
Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166.
Rogan, J. C., Keselman, H. J., & Breen, L. J. (1977). Assumption violations and rates of Type I error for the Tukey multiple comparison test: A review and empirical investigation via a coefficient of variance variation. *Journal of Experimental Education, 46*, 20–25.
Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*, 19–30.
Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*, 90–99.
Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology, 61*, 165–170.
Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* New York, NY: Springer.
Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.
Wilcox, R. R., Charlin, V., & 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*, 933–944.
Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8*, 254–274.

Received April 12, 2010
Accepted August 2, 2010
Published online August 5, 2011

John Ruscio

Psychology Department
The College of New Jersey
PO Box 7718
Ewing, NJ 08628
USA
Tel. +1 609 771-2919
Fax +1 609 637-5178
E-mail ruscio@tcnj.edu

