Psicológica (2009), 30, 371-390.

Hazards in Choosing Between Pooled and Separate-Variances *t* Tests

Donald W. Zimmerman[1] & Bruno D. Zumbo[2]
[1]Carleton University, Canada; [2]University of British Columbia, Canada

If the variances of two treatment groups are heterogeneous and, at the same time, sample sizes are unequal, the Type I error probabilities of the pooled-variances Student *t* test are modified extensively. It is known that the separate-variances tests introduced by Welch and others overcome this problem in many cases and restore the probability to the nominal significance level. In practice, however, it is not always apparent from sample data whether or not the homogeneity assumption is valid at the population level, and this uncertainty complicates the choice of an appropriate significance test. The present study quantifies the extent to which correct and incorrect decisions occur under various conditions. Furthermore, in using statistical packages, such as SPSS, in which both pooled-variances and separate-variances *t* tests are available, there is a temptation to perform both versions and to reject H_0 if either of the two test statistics exceeds its critical value. The present simulations reveal that this procedure leads to incorrect statistical decisions with high probability.

It is well known that the two-sample Student *t* test depends on an assumption of equal variances in treatment groups, or homogeneity of variance, as it is known. It is also recognized that violation of this assumption is especially serious when sample sizes are unequal (Hsu, 1938; Scheffe′, 1959, 1970). The *t* and *F* tests, which are robust under some violations of assumptions (Boneau, 1960), are decidedly not robust when heterogeneity of variance is combined with unequal sample sizes.

A spurious increase or decrease of Type I and Type II error probabilities occurs when the variances of samples of different sizes are pooled to obtain an error term. If the larger sample is taken from a population with larger variance, then the pooled error term is inflated, resulting in a smaller value of the *t* statistic and a depressed Type I error

[1] Correspondence: Donald W. Zimmerman, Ph.D. Professor Emeritus of Psychology, Carleton University. Ottawa, Canada. Mailing address: 1978 134A Street. Surrey, BC, Canada, V4A 6B6.

[2] Bruno D. Zumbo, Ph.D. Professor of Measurement and Statistics. University of British Columbia. Vancouver, BC, Canada. Email: bruno.zumbo@ubc.ca

probability. If the smaller sample is taken from a population with larger variance, the reverse is true (Overall, Atlas, & Gibson, 1995a, 1995b). See also the informative graphs presented by Hopkins, Glass, & Hopkins (1987, p. 168).

In the literature, the need to modify the *t* test when the assumption of equal variances is violated has been known as the Behrens-Fisher problem (Behrens, 1929; Fisher, 1935). Early investigations showed that the problem can be overcome by substituting separate-variances tests, such as the ones introduced by Welch (1938, 1947), and Satterthwaite (1946), for the Student *t* test, and similar methods are available for the ANOVA *F* test. These modified significance tests, unlike the usual two-sample Student *t* test, do not pool variances in computation of an error term. Moreover, they alter the degrees of freedom by a function that depends on sample data. It has been found that these procedures in many cases restore Type I error probabilities to the nominal significance level and also counteract increases or decreases of Type II error probabilities (see, for example, Overall, Atlas, & Gibson, 1995a, 1995b; Zimmerman, 2004; Zimmerman & Zumbo, 1993). In recent years, separate-variances *t* tests have become more widely used and are included in various statistical software packages such as SPSS.

The purpose of this paper is to call attention to some unforeseen dangers of this approach. For one thing, it is not easy to discern from sample data whether or not the populations from which the samples are drawn do in fact support or violate the homogeneity assumption. If a decision as to which test to employ is based on anomalous sample data that does not represent the population, the result can be misleading. Apparently homogeneous samples can arise frequently from heterogeneous populations, and vice versa. In those cases, choice of an inappropriate test can lead to a wrong decision, and sampling variability may produce that outcome more often than might be suspected.

Another difficulty arises because some researchers using software packages, such as SPSS, may be inclined to employ both pooled-variances and separate-variances tests when in doubt about the homogeneity assumption and report the result of whichever version yields statistical significance. The spurious increase in the probability of rejecting H_0 resulting from applying multiple significance tests to the same data and then picking and choosing the desired result is well established. However, the problem is somewhat more serious and less apparent in the case of the two versions of the *t* test, and incorrect statistical decisions, we shall see, can occur with very high frequency when the two tests are performed together.

The present study quantifies the extent of correct and incorrect decisions that can be expected under various scenarios where sample data is not representative of the population. It also examines the modifications of significance levels that occur when a decision is based on the Student *t* test alone, on the Welch *t* test alone, or on a favorable outcome of either test.

METHOD*

The simulations in this study were programmed using *Mathematica*, version 4.1, together with *Mathematica* statistical add-on packages. Each replication of the sampling procedure obtained two independent samples of n_1 and n_2 scores. For successive pairs, all scores in one sample were multiplied by a constant, so that the ratio σ_1/σ_2 had a predetermined value. This ratio varied from 1.0 to 1.8 in increments of .1 and then from 2 to 3 in increments of .5. In another part of the study, the ratio was 1.02, 1.05, 1.10, and 1.15. In most cases the sample sizes n_1 and n_2 varied between 10 and 60, in such a way that the ratio n_1/n_2 was 1/6, 1/5, 1/4, 1/3, 1/2, 2/3, 3/2, 2, 3, 4, 5, or 6. The total sample size $n_1 + n_2$ was fixed at 60 for the data in Tables 1 and 2 and in Figures 1, 2, and 3. There were 50,000 replications of the sampling procedure for each condition in the study.

As a check, some simulations in Figures 1 and 2 in the study were repeated using random numbers generated by the method of Marsaglia, Zaman, & Tsang (1990), described by Pashley (1993), together with normal deviates obtained by the procedure of Marsaglia & Bray (1964). These methods reproduced the results in the tables very closely, so all subsequent simulations employed random normal deviates obtained directly from the *Mathematica* statistical add-on package.

On each replication, the two-sample Student *t* test based on pooled variances, as well as the Welch-Satterthwaite version of the *t* test based on separate variances (Welch, 1938, 1947; Satterthwaite, 1946) were performed on each pair of samples, and the results were evaluated at the .01 and .05 significance levels. For the Student *t* test, the error term in the denominator of the *t* statistic was

$$\sqrt{\left[\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right]\left[\frac{1}{n_1}+\frac{1}{n_2}\right]},$$

* Listings of the *Mathematica* programs used in this study can be obtained by writing to the authors.

where n₁ and n₂ are sample sizes, and s_1^2 and s_2^2 are sample variances, evaluated at n₁ + n₂ – 2 degrees of freedom. In the case of the Welch-Satterthwaite separate-variances test, the error term was the unpooled estimate

$$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},$$

with degrees of freedom given by

$$\frac{\left[\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1-1}+\frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2-1}},$$

taken to the nearest integer.

The frequency with which each of the two test statistics exceeded its critical value was obtained over all replications to give an estimate of the probability of a Type I error for each test. In some cases, the frequencies with which either test statistic (or both) exceeded its critical value were obtained to give an estimate of the probability of a Type I error that would occur if a researcher computed both statistics and selected the most favorable one.

RESULTS OF SIMULATIONS

The results of the simulations followed essentially the same pattern for all differences in sample sizes and both significance levels. Figure 1 shows typical examples of the trend for the .05 significance level. The ratio of population standard deviations, σ_1/σ_2 varied from 1.0 to 1.8, and the sample sizes in this case were n₁ = 60 and n₂ = 10 (upper section) or n₁ = 10 and n₂ = 60 (lower section). The curves in the figure for the Student *t* test and the Welch *t* test are consistent with earlier findings (See, for example, Hopkins, Glass, & Hopkins, 1987; Hsu, 1938; Scheffe′, 1959, 1970; Zimmerman & Zumbo, 1993). When the larger variance was associated with the larger sample size (upper section), the probability of a Type I error of the Student *t* test progressively declined below .05 as the ratio of standard deviations increased. When the larger variance was associated with the smaller sample size (lower section), the probability of a Type I error increased far above .05, reaching almost .19 when the ratio was 1.8. In both cases, the separate-variances test restored the probability to a value quite close to the .05 significance level.

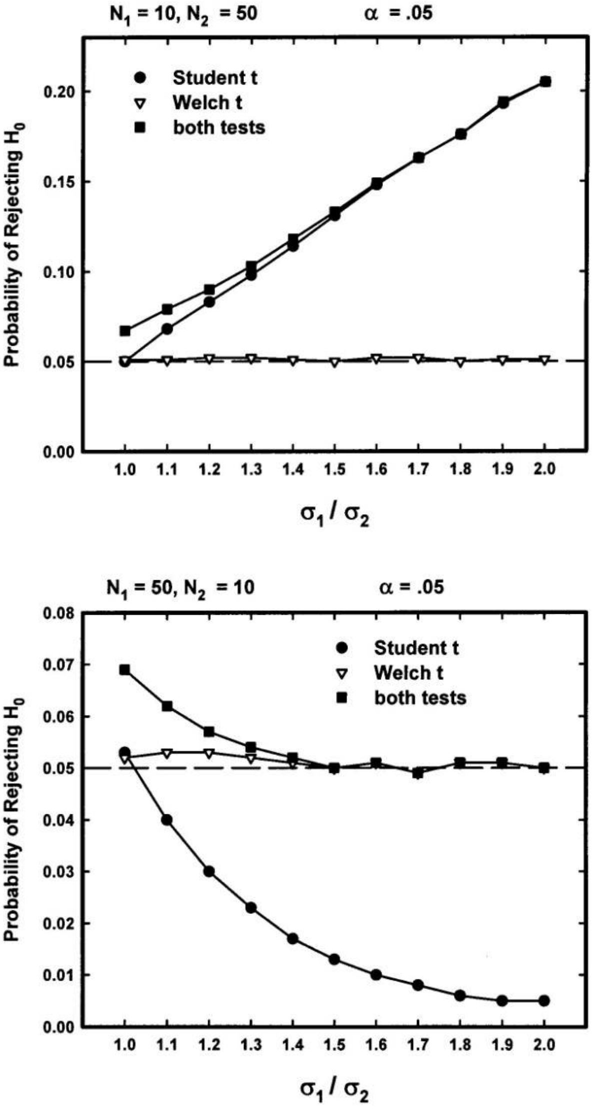


Figure 1. **Probability of rejecting H$_0$ as a function of the ratio of population standard deviations. Upper section: Larger sample is from population with smaller standard deviation. Lower section: Larger sample is from population with larger standard deviation.**

For the procedure in which the outcome of both tests determined the decision, the probability of a Type I error again depended on whether the larger variance was associated with the larger or smaller sample size. In the former case, this probability spuriously increased above the .05 significance level, although the probability for the Student *t* test alone declined below .05.

In the case where the larger variance was associated with the smaller sample size, the spurious increase in probability for the combination of both tests exceeded the increase for the Student *t* test alone but became gradually less as the ratio of standard deviations increased. Obviously, when the larger variance is associated with the smaller sample size, using both the pooled-variance test and the separate-variance test together and selecting the most favorable outcome, does not restore the .05 significance level. In fact, it makes the spurious increase in the probability of rejecting H_0 somewhat worse than it is when using the Student t test alone. Again, the change was greatest when the difference between the two standard deviations was slight rather than extreme.

Figure 2 plots the probability of rejecting H_0 as a function of n_1 when the total sample size, $n_1 + n_2$, remained constant at 60. In other words, the ratio of sample sizes, n_1/n_2 was 1/5, 1/3, ½, 5/7, 1, 7/5, 2, 3, and 5. The graphs indicate that, for the Student *t* test alone, the probability declines from about .09 to about .03, as n_1 increases. that, for the Welch *t* test alone, it remains close to .05, and that, for the combination of both tests, it declines from .09 to .05 as the two sample sizes become more nearly equal and then increases slightly as the sample sizes diverge in the opposite direction.

More complete data for a range of differences in sample sizes and for significance levels of .01 and .05 is provided by Table 1. In all cases, the results in the table follow the same trend as the results plotted in Figures 1 and 2. From these figures and the results in Table 1, it is clear why the probability of a Type I error is modified when both pooled-variances and separate-variances tests are performed. When the ratio σ_1/σ_2 is close to 1.0, each test rejects H_0 on some occasions where the other does not, because the two test statistics utilize somewhat different information. This means that the probability of rejecting H_0 always increases to some degree when H_0 is rejected if either test statistic exceeds its critical value.

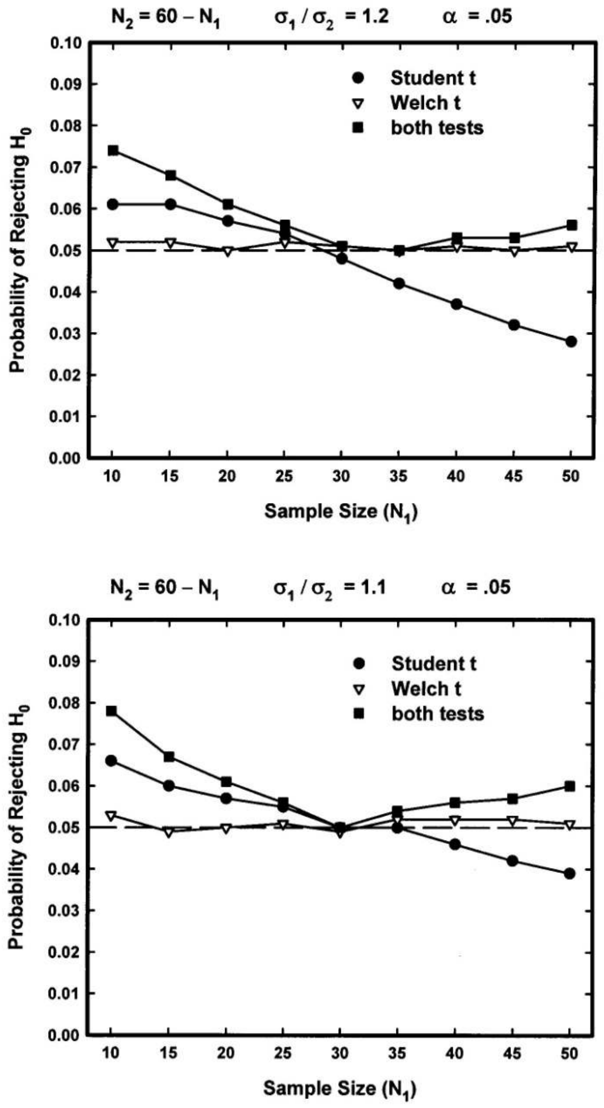


Figure 2. **Probability of rejecting H_0 as a function of the ratio of sample sizes. Upper section Ratio of population standard deviations fixed at 1.2. Lower section: Ratio of population standard deviations fixed at 3.0.**

Table 1.

Probability of rejecting H_0 by pooled-variances (Student t) and separate-variances (Welch t) tests. In condition labeled "both tests," H_0 was rejected if either pooled-variances or separate-variances test statistic exceeded its critical value.

		$n_1 = 10, n_2 = 50$			$n_1 = 50, n_2 = 10$		
σ_1/σ_2	α	Student *t*	Welch *t*	both tests	Student *t*	Welch *t*	both tests
1.0	.01	.011	.011	.016	.010	.011	.016
	.05	.052	.052	.069	.050	.050	.066
1.1	.01	.016	.011	.021	.007	.011	.014
	.05	.067	.053	.078	.036	.048	.057
1.2	.01	.024	.011	.027	.004	.011	.012
	.05	.085	.052	.093	.027	.049	.054
1.3	.01	.031	.012	.034	.003	.012	.012
	.05	.101	.050	.106	.021	.050	.053
1.4	.01	.042	.012	.044	.003	.012	.012
	.05	.121	.054	.124	.021	.050	.053
1.5	.01	.048	.011	.050	.001	.012	.012
	.05	.134	.051	.137	.013	.051	.052
2.0	.01	.097	.011	.097	.000	.012	.012
	.05	.208	.051	.208	.004	.051	.051
2.5	.01	.139	.011	.139	.000	.011	.011
	.05	.262	.053	.262	.002	.049	.049
3.0	.01	.164	.010	.164	.000	.012	.012
	.05	.290	.050	.291	.001	.049	.049

Table 1 (continued)

		$n_1 = 15, n_2 = 45$			$n_1 = 45, n_2 = 15$		
σ_1/σ_2	α	Student *t*	Welch *t*	both tests	Student *t*	Welch *t*	both tests
1.0	.01	.011	.011	.014	.011	.012	.015
	.05	.052	.050	.061	.052	.051	.062
1.1	.01	.016	.011	.018	.007	.010	.012
	.05	.064	.051	.070	.039	.050	.054
1.2	.01	.020	.011	.022	.005	.010	.011
	.05	.075	.050	.078	.033	.051	.053
1.3	.01	.025	.011	.026	.004	.011	.012
	.05	.085	.050	.088	.027	.052	.053
1.4	.01	.031	.010	.031	.003	.011	.011
	.05	.099	.050	.100	.022	.050	.051
1.5	.01	.036	.011	.037	.003	.010	.010
	.05	.109	.051	.109	.018	.048	.048
2.0	.01	.062	.011	.062	.001	.011	.011
	.05	.156	.051	.156	.010	.050	.050
2.5	.01	.084	.010	.084	.001	.011	.011
	.05	.187	.050	.187	.007	.052	.052
3.0	.01	.101	.011	.101	.001	.011	.011
	.05	.213	.051	.213	.005	.050	.050

		$n_1 = 20, n_2 = 40$			$n_1 = 40, n_2 = 20$		
1.0	.01	.014	.011	.014	.010	.011	.013
	.05	.052	.051	.058	.053	.053	.059
1.1	.01	.015	.012	.017	.009	.011	.012
	.05	.061	.051	.064	.043	.051	.053
1.2	.01	.019	.012	.019	.006	.010	.011
	.05	.067	.051	.068	.038	.050	.052
1.3	.01	.020	.012	.020	.006	.011	.011
	.05	.073	.050	.074	.033	.051	.052
1.4	.01	.021	.011	.022	.005	.012	.012
	.05	.077	.048	.077	.031	.049	.050
1.5	.01	.026	.011	.026	.005	.011	.012
	.05	.087	.050	.087	.027	.051	.051
2.0	.01	.038	.011	.038	.003	.011	.011
	.05	.108	.048	.108	.019	.051	.051
2.5	.01	.049	.011	.049	.002	.012	.012
	.05	.132	.051	.132	.015	.052	.052
3.0	.01	.056	.011	.056	.002	.012	.012
	.05	.140	.050	.140	.013	.052	.052

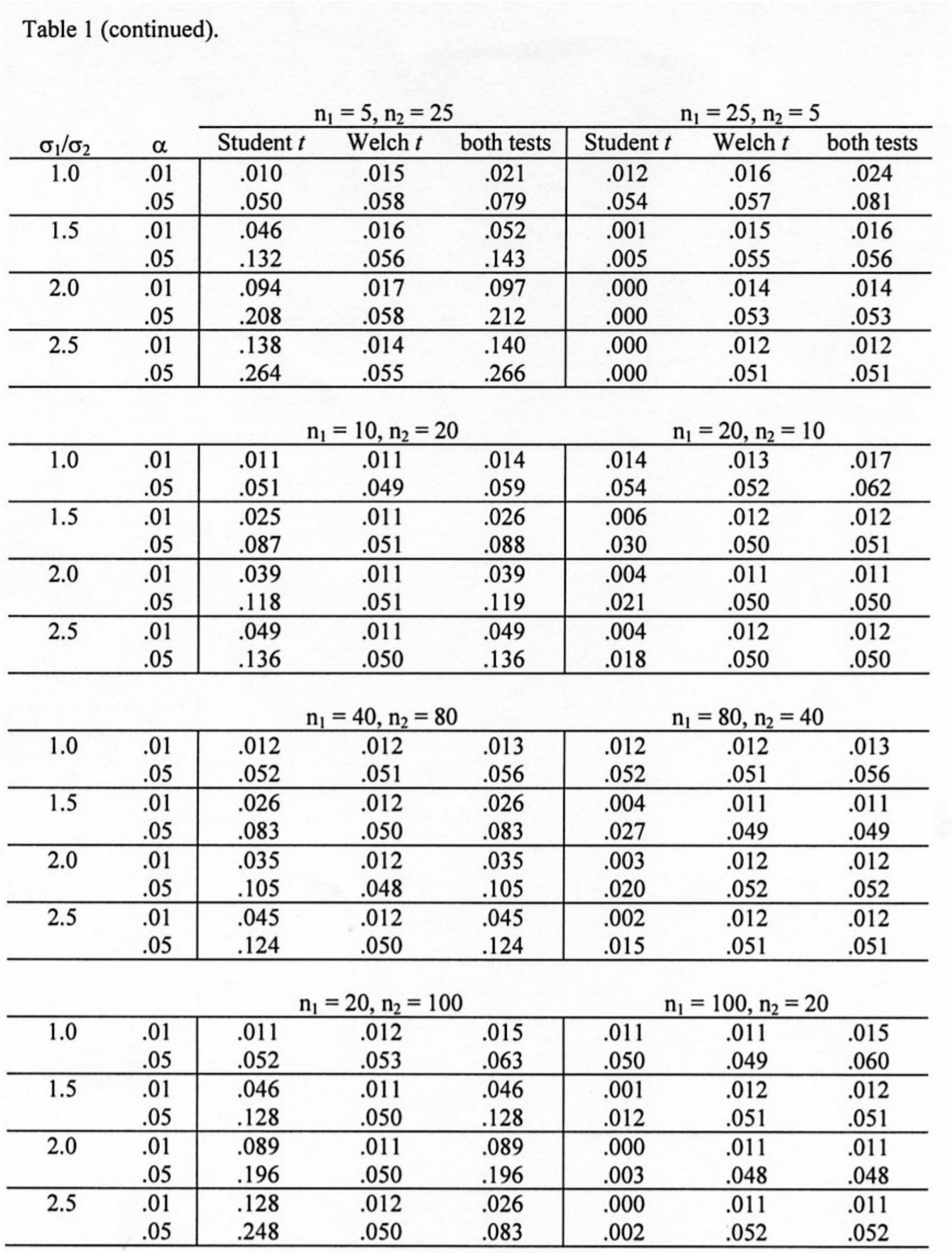


Table 1 (continued).

		$n_1 = 5, n_2 = 25$			$n_1 = 25, n_2 = 5$		
σ_1/σ_2	α	Student *t*	Welch *t*	both tests	Student *t*	Welch *t*	both tests
1.0	.01	.010	.015	.021	.012	.016	.024
	.05	.050	.058	.079	.054	.057	.081
1.5	.01	.046	.016	.052	.001	.015	.016
	.05	.132	.056	.143	.005	.055	.056
2.0	.01	.094	.017	.097	.000	.014	.014
	.05	.208	.058	.212	.000	.053	.053
2.5	.01	.138	.014	.140	.000	.012	.012
	.05	.264	.055	.266	.000	.051	.051
		$n_1 = 10, n_2 = 20$			$n_1 = 20, n_2 = 10$		
1.0	.01	.011	.011	.014	.014	.013	.017
	.05	.051	.049	.059	.054	.052	.062
1.5	.01	.025	.011	.026	.006	.012	.012
	.05	.087	.051	.088	.030	.050	.051
2.0	.01	.039	.011	.039	.004	.011	.011
	.05	.118	.051	.119	.021	.050	.050
2.5	.01	.049	.011	.049	.004	.012	.012
	.05	.136	.050	.136	.018	.050	.050
		$n_1 = 40, n_2 = 80$			$n_1 = 80, n_2 = 40$		
1.0	.01	.012	.012	.013	.012	.012	.013
	.05	.052	.051	.056	.052	.051	.056
1.5	.01	.026	.012	.026	.004	.011	.011
	.05	.083	.050	.083	.027	.049	.049
2.0	.01	.035	.012	.035	.003	.012	.012
	.05	.105	.048	.105	.020	.052	.052
2.5	.01	.045	.012	.045	.002	.012	.012
	.05	.124	.050	.124	.015	.051	.051
		$n_1 = 20, n_2 = 100$			$n_1 = 100, n_2 = 20$		
1.0	.01	.011	.012	.015	.011	.011	.015
	.05	.052	.053	.063	.050	.049	.060
1.5	.01	.046	.011	.046	.001	.012	.012
	.05	.128	.050	.128	.012	.051	.051
2.0	.01	.089	.011	.089	.000	.011	.011
	.05	.196	.050	.196	.003	.048	.048
2.5	.01	.128	.012	.026	.000	.011	.011
	.05	.248	.050	.083	.002	.052	.052

On the other hand, when the ratio σ_1/σ_2 is more extreme, one test dominates, depending on whether the larger variance is associated with the larger or smaller sample size. In the former case, the Welch test exceeds its critical value with probability close to .05, and there are relatively few cases where the Student *t* statistic exceeds its critical value and the Welch statistic does not. Substantial increases above .05 are limited to values of the ratio not too far from 1.0, for which each test rejects H_0 on a considerable number of occasions. However, when the larger variance is associated with the smaller sample size, there are substantial increases in probability of rejecting H_0 over the entire range of the ratio σ_1/σ_2. Unfortunately, reaching a decision by considering both test statistics eliminates the improvement that results from using the Welch test alone.

Table 2 presents a breakdown of the total number of sampling occasions into instances where both tests resulted in the same statistical decision, either rejection or no rejection, as well as instances where the two tests resulted in different decisions, either rejection by the Student *t* test and no rejection by the Welch test, or vice versa. Also shown are the overall probabilities of rejection by each significance test. The sum of the values in the last 4 columns always is 1.0, within the limits of rounding error. Also, the value in the column labeled “Student t test rejects” is the sum of the value in the column labeled “Both tests reject” and the value in the column labeled “Student +, Welch −.” Similarly, the value in the column labeled “Welch test rejects” is the sum of the value in the column labeled “Both tests reject” and the value in the column labeled “Student − , Welch +.”

When sample sizes are equal, there are relatively few occasions on which one test rejects H_0 and the other does not. Both tests reject H_0 with probability close to the nominal significance level, irrespective of equality or inequality of population variances. The same is true when population variances are equal but sample sizes are not. When both population variances and sample sizes are unequal and the smaller sample size is associated with the larger variance, there are no occasions on which the Welch test rejects and the Student *t* test does not. When both variances and sample sizes are unequal and the larger sample size is associated with the larger variance, the outcome is the reverse: There are no occasions on which the Student *t* test rejects and the Welch test does not. In other words, extensive inflation or depression of the probability of rejection of H_0 by the Student *t* test is associated with a difference in the outcomes of the respective tests. The two tests are almost but not quite completely consistent when sample sizes are equal or when population variances are equal.

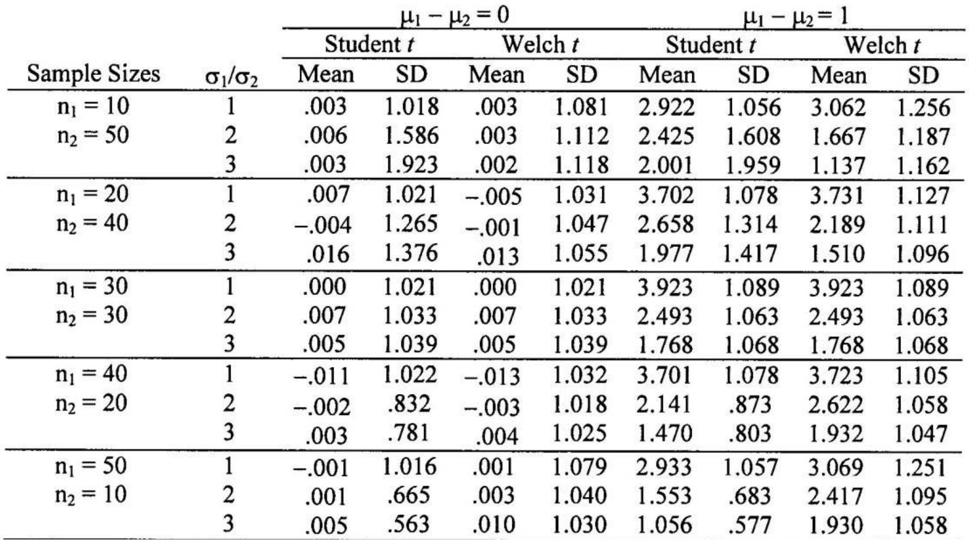
Table 2.

Probability of rejecting H_0 by Student t test, Welch t test, both tests, neither test, or one (+) but not the other (–).

			Probability					
Sample Sizes	σ_1/σ_2	α	Student *t* test rejects	Welch test rejects	Both tests reject	Neither test rejects	Student + Welch –	Student – Welch +
$n_1 = 10$ $n_2 = 50$	1	.01	.011	.013	.006	.982	.005	.007
		.05	.052	.053	.036	.930	.016	.018
	2	.01	.091	.011	.011	.908	.080	.000
		.05	.200	.051	.050	.799	.150	.000
	3	.01	.164	.011	.011	.836	.153	.000
		.05	.292	.054	.054	.708	.238	.000
	4	.01	.206	.011	.011	.794	.196	.000
		.05	.336	.051	.051	.664	.285	.000
$n_1 = 30$ $n_2 = 30$	1	.01	.012	.011	.011	.988	.000	.000
		.05	.051	.050	.050	.949	.000	.000
	2	.01	.013	.012	.012	.987	.001	.000
		.05	.054	.052	.052	.946	.002	.000
	3	.01	.014	.013	.013	.986	.002	.000
		.05	.053	.050	.050	.947	.004	.000
	4	.01	.014	.013	.013	.986	.002	.000
		.05	.053	.049	.049	.947	.004	.000
$n_1 = 50$ $n_2 = 10$	1	.01	.010	.012	.005	.984	.005	.006
		.05	.050	.051	.035	.933	.016	.016
	2	.01	.000	.012	.988	.000	.000	.012
		.05	.004	.052	.004	.948	.000	.047
	3	.01	.000	.011	.000	.989	.000	.011
		.05	.001	.050	.001	.950	.000	.049
	4	.01	.000	.012	.000	.988	.000	.012
		.05	.001	.050	.001	.950	.000	.050

Table 3.

Means and standard deviations of discrepancies between test statistics and critical values.



		$\mu_1 - \mu_2 = 0$				$\mu_1 - \mu_2 = 1$			
		Student *t*		Welch *t*		Student *t*		Welch *t*	
Sample Sizes	σ_1/σ_2	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$n_1 = 10$	1	.003	1.018	.003	1.081	2.922	1.056	3.062	1.256
$n_2 = 50$	2	.006	1.586	.003	1.112	2.425	1.608	1.667	1.187
	3	.003	1.923	.002	1.118	2.001	1.959	1.137	1.162
$n_1 = 20$	1	.007	1.021	−.005	1.031	3.702	1.078	3.731	1.127
$n_2 = 40$	2	−.004	1.265	−.001	1.047	2.658	1.314	2.189	1.111
	3	.016	1.376	.013	1.055	1.977	1.417	1.510	1.096
$n_1 = 30$	1	.000	1.021	.000	1.021	3.923	1.089	3.923	1.089
$n_2 = 30$	2	.007	1.033	.007	1.033	2.493	1.063	2.493	1.063
	3	.005	1.039	.005	1.039	1.768	1.068	1.768	1.068
$n_1 = 40$	1	−.011	1.022	−.013	1.032	3.701	1.078	3.723	1.105
$n_2 = 20$	2	−.002	.832	−.003	1.018	2.141	.873	2.622	1.058
	3	.003	.781	.004	1.025	1.470	.803	1.932	1.047
$n_1 = 50$	1	−.001	1.016	.001	1.079	2.933	1.057	3.069	1.251
$n_2 = 10$	2	.001	.665	.003	1.040	1.553	.683	2.417	1.095
	3	.005	.563	.010	1.030	1.056	.577	1.930	1.058

Sampling Variability of Heterogeneity of Variance

Because of the inescapable presence of random sampling in significance testing of differences between means, it is possible to obtain pairs of heterogeneous samples from homogeneous populations, as well as pairs of homogeneous samples from heterogeneous populations. For this reason, a decision to use the pooled or separate variances *t* test based on the data obtained from two samples, s_1^2 and s_2^2, may not be consistent with a decision based on the known population parameters, σ_1^2 and σ_2^2.

This state of affairs probably occurs more often than one might suspect. Some examples that could occur in practical research are shown in Figure 3, based on 50,000 replications of the sampling procedure. The top section of the figure is a frequency distribution of ratios of the larger to the smaller sample standard deviations, when samples of size 8 and 12, respectively, are taken from two populations in which the variances are equal. Obviously, the sample ratios vary over a wide range and include many values that would suggest a violation of the homogeneity assumption. The reverse situation is shown in the lower section. In this case, the ratio of population standard deviations is 1.5, which indicates the need for a separate-variances *t* test, especially since the sample sizes are unequal. In both cases, the ratios are clustered around the population values, but the tails of the distributions include many ratios far from those values. The mean of the sample ratios in the upper section is 1.041, and the standard deviation is .391. The mean of the ratios in the lower section is 1.556, and the standard deviation is .581.

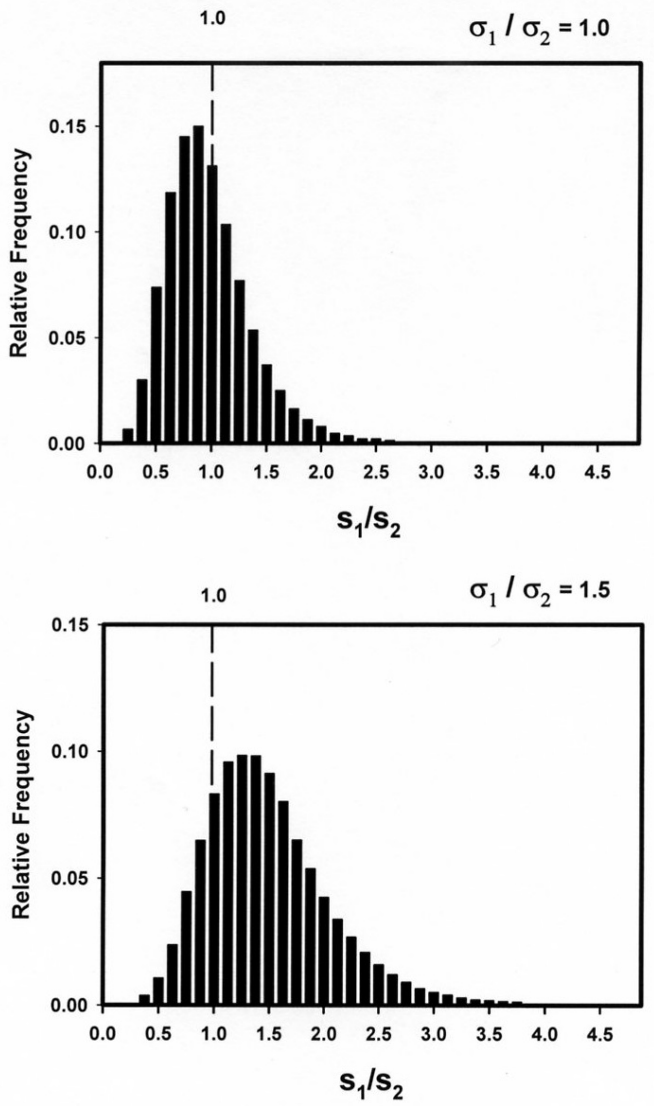


Figure 3. **Relative frequency distributions of ratios of sample standard deviations when ratios of population standard deviations are 1.0 (upper section) and 1.5 (lower section).**

Table 4 provides more specific information about these uncharacteristic outcomes. The values in the table are proportions of instances in which the sample ratios exceed various cutoff values that might be taken to indicate homogeneity or heterogeneity. The left-hand section, in which the population standard deviations are equal, includes cutoff values at successively higher ratios of sample standard deviations that could be taken to indicate heterogeneity. The right-hand section, where the population ratio is 1.5, includes successively lower ratios that could mistakenly indicate homogeneity. When sample sizes are relatively small, up to about 20, the proportions of anomalous samples are quite large. However, as sample sizes increase to 50, 100, or higher, the proportions decline and eventually become close to zero. It is clear that inappropriate selection of the significance test under this scenario is largely restricted to small-sample research settings. For the very small sample sizes, the decision can be wrong in a substantial proportion of cases.

Table 4.

Proportion of samples in which variances appear heterogeneous (s_1/s_2 or $s_2/s_1 > k$) when populations are homogeneous ($\sigma_1/\sigma_2 = 1$) and proportion of samples in which variances appear homogeneous (s_1/s_2 and $s_2/s_1 < k$) when populations are heterogeneous ($\sigma_1/\sigma_2 = 1.5$).

	$\sigma_1/\sigma_2 = 1$				$\sigma_1/\sigma_2 = 1.5$			
	> cutoff ratio of sample SDs (k)				< cutoff ratio of sample SDs (k)			
Sample Sizes	1.1	1.25	1.5	2	1.2	1.15	1.1	1.05
$n_1 = 4, n_2 = 6$	.861	.688	.468	.230	.198	.150	.103	.052
$n_1 = 5, n_2 = 5$	.857	.672	.446	.210	.194	.151	.106	.053
$n_1 = 6, n_2 = 9$	.822	.594	.341	.114	.219	.170	.115	.058
$n_1 = 10, n_2 = 5$	.831	.619	.368	.138	.203	.152	.107	.053
$n_1 = 8, n_2 = 12$	.785	.530	.260	.060	.223	.170	.113	.058
$n_1 = n_2 = 10$	.777	.517	.246	.054	.212	.164	.109	.056
$n_1 = n_2 = 20$	.679	.335	.084	.004	.161	.117	.073	.036
$n_1 = n_2 = 40$	.552	.163	.012	.000	.084	.049	.025	.011
$n_1 = n_2 = 80$	.394	.047	.000	.000	.023	.009	.003	.001
$n_1 = n_2 = 160$	.228	.005	.000	.000	.002	.000	.000	.000
$n_1 = n_2 = 320$	.087	.000	.000	.000	.000	.000	.000	.000

Effects of Slight Variance Heterogeneity Together with Unequal Sample Sizes

The probability of rejecting H_0 by the Student *t* test is severely inflated or depressed when sample sizes are unequal at the same time variances are heterogeneous. If one of the two is not an issue— that is, if variances are equal or if sample sizes are the same—the influence of the other is not large. However, when both conditions occur at the same time, the influence of one can be greatly magnified by changes in the other.

Table 5 exhibits some cases where ratios of population standard deviations fairly close to 1.0 produce large changes in the probability of rejecting H_0 provided sample sizes differ greatly. For example, a ratio of standard deviations of only 1.05 elevates or depresses the probability of rejecting H_0 considerably above or below the nominal significance level when the difference in sample sizes is large, such as $n_1 = 10$ and $n_2 = 100$. Even a ratio of only 1.02 has noticeable effects under these large differences in sample sizes. The data in Table 5 also show that the Welch test still restores the probability to a value fairly close to the nominal significance level in these extreme cases.

Considering the results in Table 5 together with the results of sampling variability shown in the foregoing tables, it becomes clear that there is a danger in disregarding slight differences in sample variances and assuming that the Student *t* test is appropriate, especially if the difference in sample sizes is large. The slight differences in standard deviations included in Table 5, we have seen, can be expected to arise frequently from sampling variability even though population variances are equal. In such instances, a large difference in sample sizes can lead to an incorrect statistical decision by the Student *t* test with high probability

FURTHER DISCUSSION

If multiple significance tests are performed on the same data at the same significance level, and if H_0 is rejected when any one of the various test statistics exceeds its critical value, the probabilities of Type I and Type II errors are certainly modified to some extent. For example, if a Student *t* test and a Wilcoxon-Mann-Whitney test are performed on the same data, at the .05 significance level, then this omnibus procedure makes the Type I error probability about .055, or perhaps .06. For other significance tests of differences between means, the outcome is about the same. In the case of most commonly used significance tests of location, the change is not large, and textbooks pay relatively little attention to this illegitimate procedure.

The result is always an increase in Type I error probability, because the critical region for multiple tests is a union of the critical regions of the individual tests. Since various significance tests utilize much the same information in the data, the change usually is minimal.

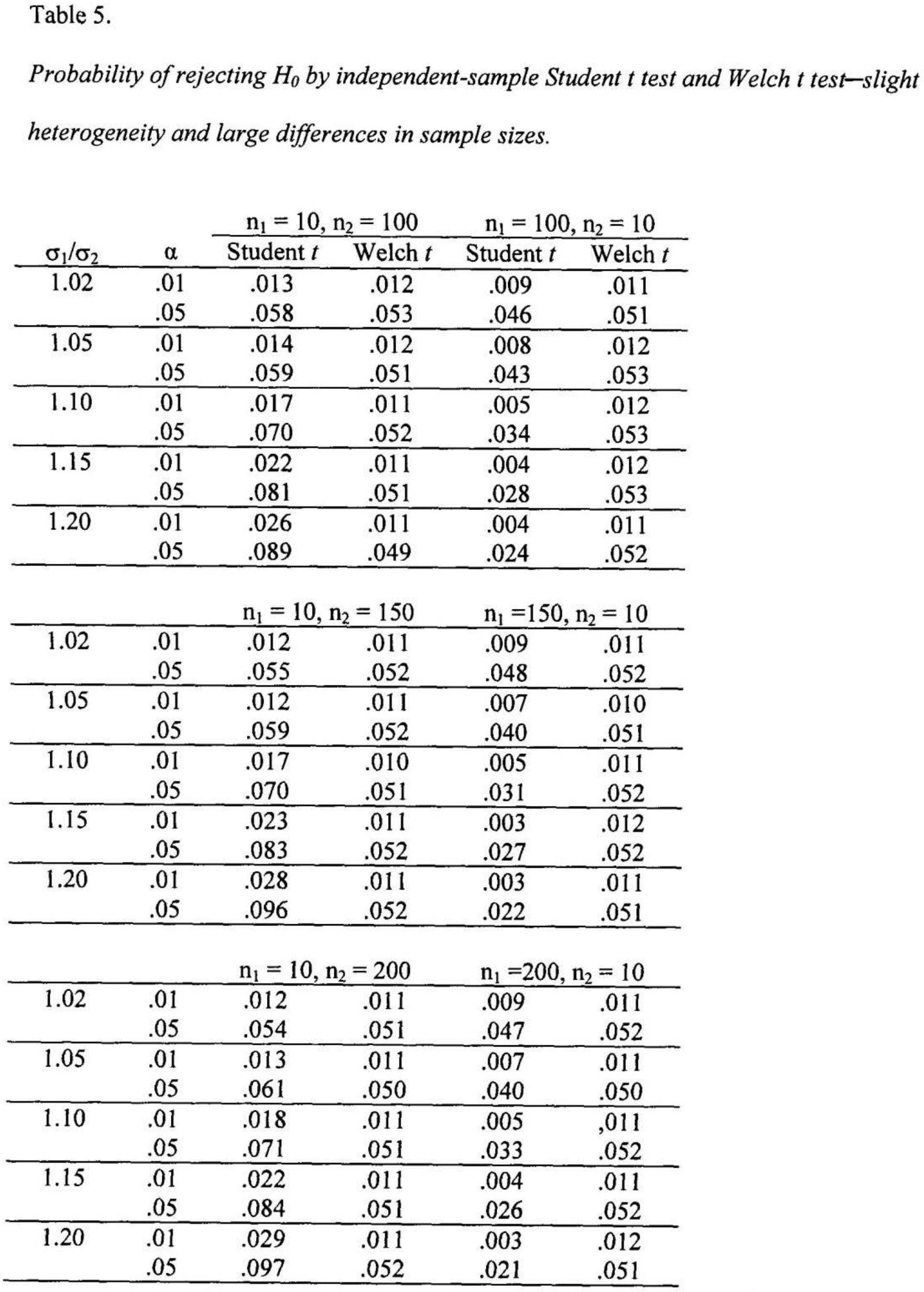


Table 5.

Probability of rejecting H_0 by independent-sample Student t test and Welch t test—slight heterogeneity and large differences in sample sizes.

		$n_1 = 10, n_2 = 100$		$n_1 = 100, n_2 = 10$	
σ_1/σ_2	α	Student *t*	Welch *t*	Student *t*	Welch *t*
1.02	.01	.013	.012	.009	.011
	.05	.058	.053	.046	.051
1.05	.01	.014	.012	.008	.012
	.05	.059	.051	.043	.053
1.10	.01	.017	.011	.005	.012
	.05	.070	.052	.034	.053
1.15	.01	.022	.011	.004	.012
	.05	.081	.051	.028	.053
1.20	.01	.026	.011	.004	.011
	.05	.089	.049	.024	.052
		$n_1 = 10, n_2 = 150$		$n_1 =150, n_2 = 10$	
1.02	.01	.012	.011	.009	.011
	.05	.055	.052	.048	.052
1.05	.01	.012	.011	.007	.010
	.05	.059	.052	.040	.051
1.10	.01	.017	.010	.005	.011
	.05	.070	.051	.031	.052
1.15	.01	.023	.011	.003	.012
	.05	.083	.052	.027	.052
1.20	.01	.028	.011	.003	.011
	.05	.096	.052	.022	.051
		$n_1 = 10, n_2 = 200$		$n_1 =200, n_2 = 10$	
1.02	.01	.012	.011	.009	.011
	.05	.054	.051	.047	.052
1.05	.01	.013	.011	.007	.011
	.05	.061	.050	.040	.050
1.10	.01	.018	.011	.005	,011
	.05	.071	.051	.033	.052
1.15	.01	.022	.011	.004	.011
	.05	.084	.051	.026	.052
1.20	.01	.029	.011	.003	.012
	.05	.097	.052	.021	.051

However, in the case of the pooled-variances and separate-variances *t* tests, the outcome of such a procedure is quite different and more serious. The disparity arises as a consequence of two effects of heterogeneous variances. First, the probability of a Type I error is altered extensively when the ratio of variances and the ratio of sample sizes are both large. Second, the direction of the change depends on whether the larger variance is associated with the larger or smaller sample size. In the first case the probability of a Type I error declines below the nominal significance level, and in the second case it increases (see Figure 1 and Table 1).

Suppose, for example, that the ratio σ_1/σ_2 is about 1.1, that $n_1 = 60$, $n_2 = 10$, and that both pooled-variances and separate-variances tests are performed at the .05 significance level. Then, it is clear from Figure 1 that rejecting H_0 when either test statistic exceeds its critical value makes the probability of a Type I error about .06 instead of .05. This can be contrasted with the decline to .035 when the pooled-variances test alone is performed. However, the probability for the separate-variances test alone is close to .05.

Next, suppose the ratio σ_1/σ_2 remains 1.1, while the sample sizes are reversed: $n_1 = 10$ and $n_2 = 60$. Then, rejecting H_0 when either test statistic exceeds its critical value makes the probability of a Type I error about .08, which is worse than the .07 resulting from the pooled-variances test alone. Again, the separate-variances test alone yields the correct value of .05. If the ratio σ_1/σ_2 is somewhat larger, say, 1.6, then performing both tests makes the probability of a Type I error about .15, which is the same as employing the pooled-variances test alone. In this case, rejecting H_0 after applying the separate-variances test alone has a more favorable outcome.

The result of performing both pooled-variances and separate-variances tests on the same data, therefore, is not simply an increase in the probability of rejecting H_0 to a value slightly above the nominal significance level, because of capitalizing on chance. That is true only if the ratio σ_1/σ_2 falls in a rather narrow range between about 1.0 and about 1.5 and, in addition, the larger sample is associated with the larger variance. In all cases of heterogeneous variances examined in the present study, employing the separate-variance test alone resulted in rejecting H_0 with probability close to the nominal significance level. In most cases, the pooled-variances test, either alone or in combination with the separate-variances test, substantially altered the probability. To look at it another way, reaching a decision by considering both the pooled and separate-variances test statistics together eliminates the improvement that comes from using the separate-variances test alone.

Another complication arises from the fact that researchers may not always know when population variances are heterogeneous. And in some cases it is even problematic whether a larger variance is associated with a larger or smaller sample size. If the ratio of sample sizes is quite different from 1.0, say 1/4 or 1/5, then the probability of a Type I error is sensitive to slight differences in population variability. For example, as we have seen, an apparently inconsequential ratio of standard deviations of 1.1 can increase the probability of a Type I error substantially. On the other hand, interchanging the larger and smaller standard deviations, so that the ratio becomes .91, significantly decreases the nominal Type I error probability.

Because sample variances do not always reflect population variances, it is difficult to determine whether Type I error probabilities will increase or decrease if sample sizes differ. A ratio of sample standard deviations of 1.1 or 1.2 could easily arise in sampling from populations with ratios considerably more extreme and possibly when the ratio is less than 1.0. The sensitivity of the Type I error probabilities to small differences, makes it risky to reach a decision about the appropriate test solely on the basis of sample statistics.

Preliminary tests of equality of variances, such as the Levene test, the *F* test, or the O'Brien test, are ineffective and actually make the situation worse (Zimmerman, 2004). That is true because, no matter what preliminary test is chosen, it inevitably produces some Type II errors. On those occasions, therefore, a pooled- variances test is performed when a separate-variances test is needed, modifying the significance level to some degree. On the other hand, a Type I error, resulting in using a separate-variances test when variances are actually homogeneous, is of no consequence.

Instead of inspecting sample data and performing a preliminary test of homogeneity of variance, it is recommended that researchers focus attention on sample sizes. Then, to protect against possible heterogeneity of variance, simply perform a separate-variances test unconditionally whenever sample sizes are unequal. That strategy protects the significance level if population variances are unequal, whatever the outcome of a preliminary test might be. And if variances are in fact homogeneous, nothing is lost, because then both pooled and separate-variances tests lead to the same statistical decision.

REFERENCES

Behrens, W.U. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftlisches Jahrbuch, 68,* 807-837.

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

Fisher, R.A. (1935). *The design of experiments.* Edinburgh: Oliver & Boyd.

Hopkins, K.D., Glass, G.V., & Hopkins, B.R. (1987). *Basic statistics for the behavioral sciences* (2nd ed.). Prentice-Hall: Englewood Cliffs, NJ.

Hsu, P.L. (1938). Contributions to the theory of Student's *t* test as applied to the problem of two samples. *Statistical Research Memoirs, 2,* 1-24.

Marsaglia, G., & Bray, T.A. (1964). A convenient method for generating normal variables. *SIAM Review, 6,* 260-264.

Marsaglia, G., Zaman, A., & Tsang, W.W. (1990). Toward a universal random number generator. *Statistics and Probability Letters, 8,* 35-39.

O'Brien, R.G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin, 89,* 570-574.

Overall, J.E., Atlas, R.S., & Gibson, J.M. (1995a). Tests that are robust against variance heterogeneity in k × 2 designs with unequal cell frequencies. *Psychological Reports, 76,* 1011-1017.

Overall, J.E., Atlas, R.S., & Gibson, J.M. (1995b). Power of a test that is robust against variance heterogeneity. *Psychological Reports, 77,* 155-159.

Pashley, P.J. (1993). On generating random sequences. In G. Keren and C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp 395-415). Hillsdale, NJ: Lawrence Erlbaum Associates.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110-114.

Scheffe′, H. (1959). *The analysis of variance.* New York: Wiley.

Scheffe′, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association, 65,* 1501-1508.

Welch, B.L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350-362.

Welch, B.L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika, 34,* 29-35.

Zimmerman, D.W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57,* 173-181.

Zimmerman, D.W., & Zumbo, B.D. (1993). Rank transformations and the power of the Student *t* test and Welch t' test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47,* 523-539.

(Manuscript received: 28 August 2008; accepted: 10 December 2008)