Bulletin of the Psychonomic Society
1984, 22 (3), 250-253

The complexity of nonrobustness effects

JAMES V. BRADLEY
New Mexico State University, Las Cruces, New Mexico

The factors determining nonrobustness are numerous and their effects are complicated, often being interactive rather than additive. This makes it hazardous to generalize the results of robustness studies beyond the exact conditions investigated in a study or to attempt to predict the unknown effects of several factors in combination on the basis of their known effects in isolation. Among other things, it is shown (with graphs based on results of a large sampling study) that when both the normality assumption and the equal-variances assumption of certain classical tests on means are violated, (1) the degree of nonrobustness resulting may far exceed the sum of the degrees resulting from the same specific violations of the normality assumption alone and the equal-variances assumption alone, (2) robustness does not necessarily increase with increasing absolute sample size or with enlarging significance levels, and (3) taking equal-sized samples does not necessarily prevent extreme nonrobustness (in fact, it does not necessarily do so even if only the normality assumption is violated).

A previously reported study of robustness (Bradley, 1980) was concerned primarily with the "slowness" with which robustness developed, and its results were given only very partially and indirectly by stating the sample sizes at which certain criteria of robustness were met. The present article is concerned with the complexity and extremity of nonrobustness effects, which can best be shown graphically. Such graphs have the additional advantage of clearly revealing the course and flow of nonrobustness effects as sample size increases; furthermore, the error of estimate can be inferred from the smoothness of the curves. The graphs to be presented are based upon data obtained in the previously reported study and represent only certain highly selected cases from the many conditions investigated in that study that might have served to illustrate the points to be made.

METHOD

The methodology has been described in detail elsewhere (Bradley, 1980). For present purposes, only the following fragmentary information is necessary. There were four populations from which samples could be drawn: (1) a highly skewed L-shaped population denoted by an X, (2) an essentially normally distributed population having the same mean and variance as the X population and denoted by a Y, and (3 and 4) two populations identical to the X and Y populations, respectively, except that their common standard deviation was exactly half as large as the common standard deviation of the X and Y populations; these last two populations are denoted by an A and a B, respectively (although in the previous study they were denoted by an x and a y, respectively). A value of the two-independent-sample t statistic was obtained by pseudorandomly drawing a sample of N_1 observations from a "first population," Population 1 (which was always either X or Y), and a sample of N_2 observations from a "second population," Population 2 (X, Y, A or B), and calculating the value of the test statistic (whose numerator is the mean of the first sample minus the mean of the second sample). The relative sizes of N_1 and N_2 could be any of the following: N and 2N, 2N and 2N (the notation here differs from that used in the previous study), or 2N and N, where N is an index of absolute sample size, which could have any of the values 2, 4, 8, 16, 32, 64, 128, 256, 512, or 1,024. A value of the independent-sample F statistic was obtained by pseudorandomly drawing a sample of N observations from Population 1 (which was always X) and either two or three samples, each consisting of N observations, from Population 2 (which was always A), where N is the above-mentioned index of absolute sample size. For each distinguishably different t statistic (i.e., for designated Population 1, Population 2, N_1, N_2, and N) or F statistic (i.e., for designated Population 1, Population 2, N, and number of samples drawn from Population 2), a sampling distribution of at least 10,000 values of the tests statistic was obtained. For each such sampling distribution, the proportion ρ of the sampling distribution falling in a rejection region for which the expected proportion would have been α if all assumptions had been met, was determined. This was done for left-tailed, two-tailed, and right-tailed rejection regions and for $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$.

The author's mailing address is: Department of Psychology, New Mexico State University, Las Cruces, New Mexico 88003.

RESULTS

The results are shown in a series of graphs (see Figures 1 and 2). At the top of each graph, a code that identifies the graph is given. The first letter, T or F, identifies the test (t or F), and the second and third letters identify Populations 1 and 2, respectively; after a space, these are followed, for the t test, by the relative sizes of the samples drawn from Populations 1 and 2, respectively, or, for the F test, by as many Ns as there were samples, a comma dividing them into those representing samples drawn from Population 1 and from Population 2, respectively. Thus, TXA N2N identifies the graphs giving results for the t test based upon N observations drawn from Population X and 2N observations drawn from Population A; FXA N,NN and FXA N,NNN identify graphs giving results for the F test based upon one sample of N observations drawn from Population X and two or three samples, respectively,

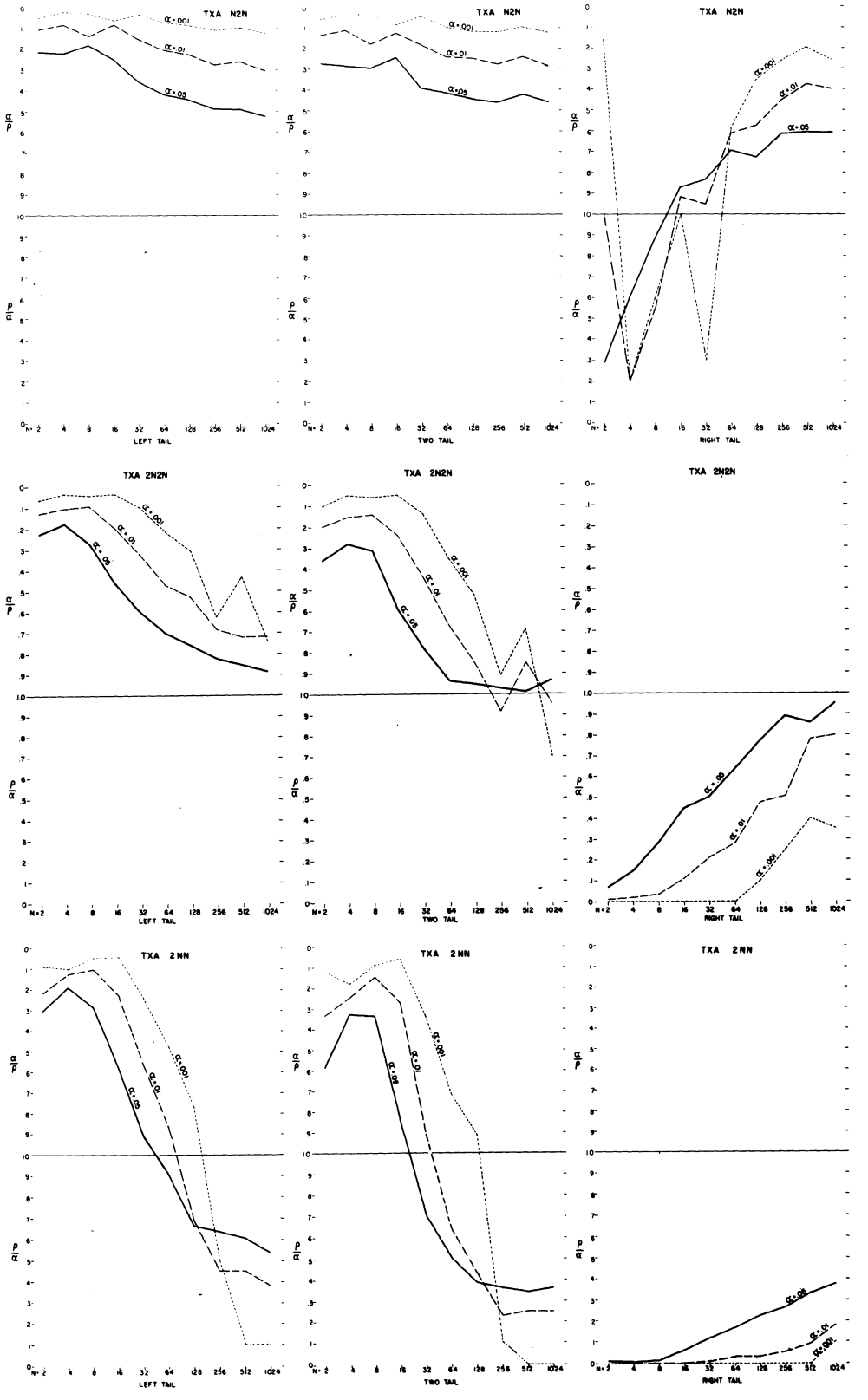

Figure 1. Changing patterns of nonrobustness of t against L-shapedness and heterogeneity with differing relative sample sizes.

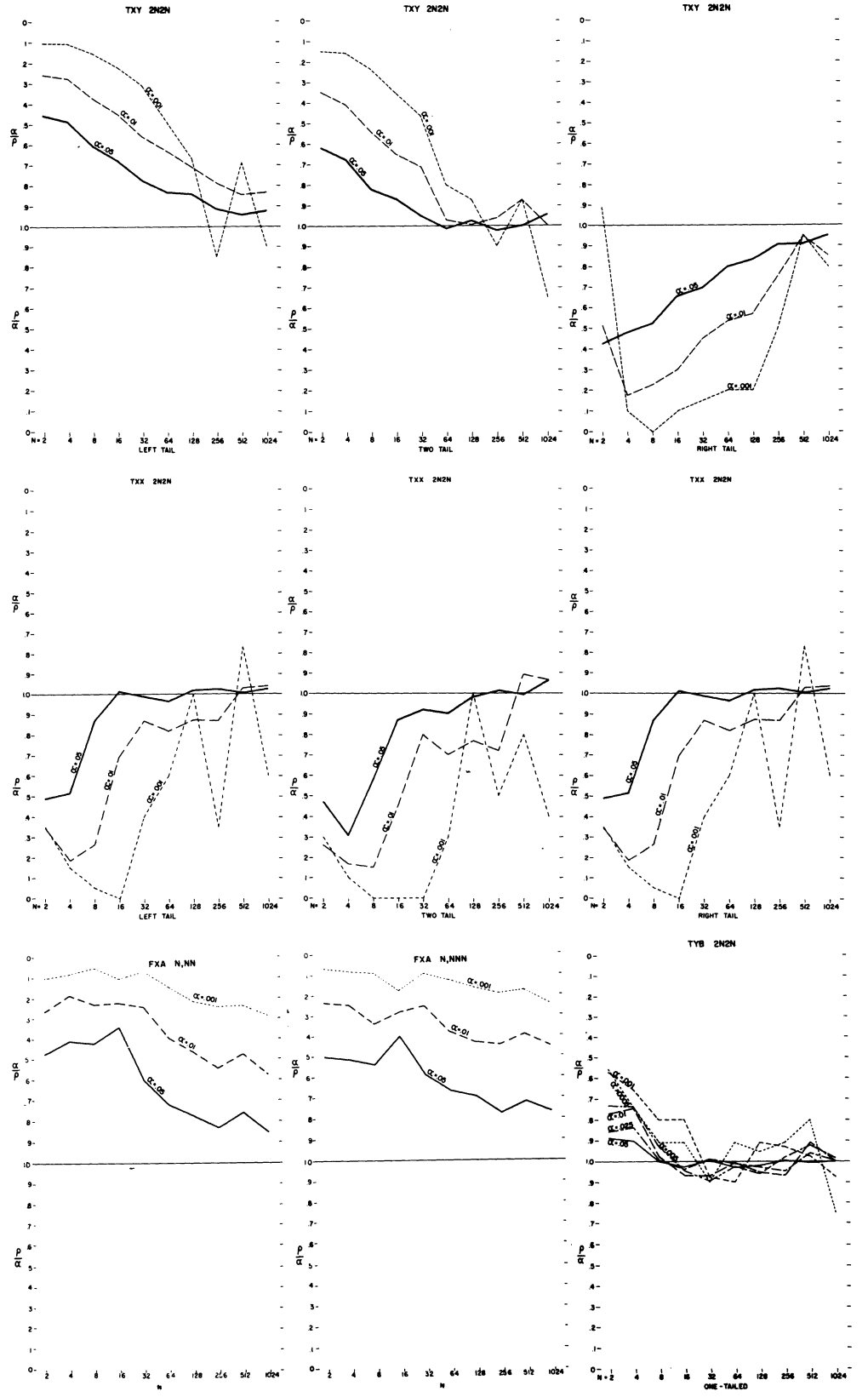

Figure 2. Nonrobustness despite equal-sized samples, for t and F in various situations; and (lower right graph) robustness of t against heterogeneity alone when sample sizes are equal.

each consisting of N observations, drawn from Population A. It is important to remember that, in the case of the t tests (unlike the notation used in Bradley, 1980) N is not necessarily the size of a sample, as when the two sample sizes are 2N and 2N (e.g., when N = 2, TXA 2N2N represents a t test based upon two samples, each containing four observations, and therefore upon a total of eight observations). Unless this is borne in mind, the "2N2N" graphs will appear to indicate much more robustness than they actually do.

For each of the 10 different possible N values (on the abscissa scale), the graphs plot (on the ordinate scale) a ratio between ρ, the actually obtained empirical probability of a Type I error, and α, the alleged probability, that would have been true if all assumptions had been met. When ρ is less than α, the ratio of ρ to α is plotted against an ascending scale in the lower half of the graph; when ρ exceeds α, the ratio of α to ρ is plotted against a descending scale in the upper half. (The result is the same as if ρ/α had been used throughout, being plotted against an ordinate scale that is linear from 0 to 1 and reciprocal from 1 to infinity.) In each graph (with some appropriate additional curves for TYB 2N2N), the ratio between ρ and α is indicated by a solid line for $\alpha = .05$, a dashed line for $\alpha = .01$, and a dotted (or shorter dashed) line for $\alpha = .001$. For all t statistics except TYB 2N2N (whose symmetric distribution makes it unnecessary), each rejection region is represented by a different graph, and for these statistics, in each tier of graphs, the leftmost graph represents the left-tailed test, the middle graph concerns the two-tailed test, and the rightmost graph deals with the right-tailed test. Thus, each plotted point on any graph is ultimately based upon the proportion of ts or Fs in a sampling distribution of at least 10,000 t or F values that fell in a normal-theory "α-sized" rejection region.

DISCUSSION

The figures show that robustness effects vary greatly with particular circumstances, such as the statistic involved, the value of α, the location of the rejection region, the particular populations sampled, absolute sample size, relative sample sizes, and which particular population contributed the sample of which relative size (compare TXA N2N with TXA 2NN), and that these effects are highly interdependent. For example, although when based upon equal-sized samples, the t test is well known to be highly robust against heterogeneity alone (and this is impressively confirmed by the graph for TYB 2N2N) and is widely regarded as robust against nonnormality alone (which the graphs for TXX 2N2N do not, in general, confirm but do moderately support when $\alpha = .05$ and $N \geq 16$), it can be highly nonrobust against the combination of these two violations (as shown by the graphs for TXA 2N2N, which can be regarded as combining the two violations individually and separately present in TXX 2N2N and TYB 2N2N). Clearly, the nonrobustness shown in the graphs for TXA 2N2N far exceeds the "sum" of the individual nonrobustness effects shown in the graphs for TXX 2N2N and TYB 2N2N (i.e., departures from 1 of the ratio between ρ and α in the former case tend greatly to exceed the sum of corresponding departures in the two latter cases). Furthermore, the degree of nonrobustness shown in the graphs for TXA 2N2N is dramatic at small and moderate sample sizes, and even when each sample contains 2,048 observations, the one-tailed test is impressively robust only at the largest of the three standard α values. Unfortunately, textbooks seldom, if ever, warn us that individually innocuous violations can produce drastic nonrobustness when combined. Thus, it appears that taking equal-sized samples is not a panacea for nonrobustness effects; not only may the panacea fail when both the normality and homogeneity assumptions are violated, but the graphs for TXY 2N2N show that it may also fail when only the normality assumption is violated. Merely sampling from populations of different shapes (only one of which is nonnormal) may produce an impressive degree of nonrobustness that dissipates only very slowly with increasing sample size.

We have seen that, when samples are of equal size, the t test may be distressingly nonrobust against the combination of nonnormality and heterogeneity at small, moderate, and even large absolute sample sizes. However, it must eventually become satisfactorily robust as absolute sample size continues to increase, since it can be shown mathematically (Bradley, 1968; Scheffé, 1959) that it becomes perfectly robust against this combination of violations when the common size of the two samples becomes infinite. Unfortunately, however, when based upon more than two samples, the F test does not share this property. When sample size becomes infinite, the equal-sample F test, based upon three or more samples, becomes perfectly robust against nonnormality but not against heterogeneity (Bradley, 1968; Scheffé, 1959). Thus, the graphs for FXA N,NN and FXA N,NNN show a disquieting degree of nonrobustness at sample sizes in common use; and even when each of the three or four samples contains 1,024 observations, the curves for α/ρ still lie appreciably (and alarmingly for αs of .01 and .001) above 1, which is definitely not their asymptote.

Not only are nonrobustness effects highly interdependent, but also the resulting complexity is of such a high order that it makes even the simplest predictions hazardous. It may be extremely difficult even to predict the "direction" of the nonrobustness, that is, whether the true probability of a Type I error (estimated by ρ) will exceed or be less than the nominal probability, α; and the seemingly safe predictions that robustness will increase with increasing absolute sample size, N, or with increasingly large significance level, α, may be badly in error. All these assertions are verified by the graph for TXA 2NN, two-tailed test, in which at N = 128 the test is far more robust for $\alpha = .001$ than for either of the two larger α values and in which robustness (as defined by the closeness to 1 of the ratio between ρ and α) at $\alpha = .05$ first decreases with increasing N values, then increases, becomes "perfect" at about N = 16, and then decreases: That is, ρ first exceeds α, then, with increasing N values, exceeds it even more, reverses direction and approaches α, passes through it becoming smaller than α, and then diminishes still further, finally becoming about a third of α when N reaches about 1,000. The "perfect" robustness at about N = 16 provides little reason to rejoice, since its causes are too complicated and the effect (crossing over from $\rho > \alpha$ to $\rho < \alpha$) too ephemeral and peculiar to circumstances (such as α values) to be predictable and therefore to be exploited.

Clearly, nonrobustness effects are extremely complex, being the result of high-order interactions among a considerable number of determining factors that are related not only to populations (e.g., population shape and relative variance), but also to sampling (absolute and relative sample size) and testing (α level and rejection region).

REFERENCES

BRADLEY, J. V. (1968). Studies in research methodology. *Dissertation Abstracts*, **28**, 4815B-4816B. (Monograph)

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

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

(Manuscript received for publication February 6, 1984.)