COMMENT ON THE RELATIVE USEFULNESS OF THE TWO INDEPENDENT MEANS _T_-TEST AND WILCOXON'S RANK-SUM TEST FOR THE ANALYSIS OF EDUCATIONAL DATA

R. C. Blair

University of South Florida

Researchers and evaluators in education are often interested in assessing the effectiveness of some treatment through a test of the hypothesis $f(x) = f(x-\theta)$, where $f(x)$ is the functional form of the population(s) about which inferences are to be made and θ is some constant that represents the treatment effect. Among educational researchers, by far the most common statistic employed for this test is the two independent means _t_-test. This latter statement is true even though no specific or limiting conditions were placed upon $f(x)$. It will be contended here that this practice should be reexamined and that Wilcoxon's rank-sum (or the equivalent Mann-Whitney U) test should be substituted for the _t_-test in most of the situations where the latter statistic is routinely employed. It will be further contended that the present state of affairs was brought about by the exaggeration of facts and, more importantly, by an unclear understanding of the involved issues.

During the 1950's, Wilcoxon's statistic along with various other nonparametric procedures gained some popularity with researchers in education and the social sciences in general. Movement away from the Wilcoxon statistic was prompted by the argument that the _t_-test is _functionally_ a distribution-free test; therefore, there is no need to revert to the less powerful nonparametric procedures (Boneau, 1960). This arguments seems to be the basis for the contention by Glass, Peckhan and Sanders (1972) that the use of nonparametric tests is "largely unnecessary" (p. 237). These authors go on to warn that "Incautious statements concerning the robustness of the ANOVA to non-normality could send applied statistics off on a return of the unproductive 1950s stampede to nonparametric methods" (p. 255). This statement was made in response to a rather mild statement by Hawkridge (1970) who suggested that nonparametric tests might usefully be substituted for parametric tests in some circumstances.

Perhaps the most flagrantly incautious statement concerning the robustness of the _t_-test to non-normality was that made by Glass and Stanley (1970) who assert that "Violation of the assumption of normality in the _t_-test of H_0: $\mu_1-\mu_2 = 0$ has been shown to have only _trivial_ [italics added] effects on the level of significance and the power of the test and hence should be no cause for concern" (p. 297). Although it is true that the _t_-test is remarkably robust to many forms of population non-normality, it is _not_ universally so. Investigations have shown that large discrepancies develop between nominal and empirical significance levels when the _t_-test is conducted on samples drawn from certain real and theoretical populations (Blair & Phillippy, 1978; Bradley, 1964, 1968). Although the determination of what constitutes "large discrepancies" and what does not must to some degree rest in the eye of the beholder, the discrepancies found in the cited studies cannot be termed "trivial" by any but the most exaggerated standards. (Interestingly, a journal referee once defended this statement by pointing out that it was meant for neophytes in the area. Justification for unnecessary exaggerations of facts on the grounds that the readers are too unsophisticated to know any better requires a form of logic that, quite frankly, escapes this author.)

Although it is true that most authors are not as incautious as Glass and Stanley (1970) in their statements concerning the robustness of the _t_-test to population non-normality, it is nevertheless true that their statements are rarely, if ever, sufficiently qualified so that they accurately reflect the known facts (Bradley, 1978). The problem then is not a lack of caution on the part of those who point to the fact that the _t_-test shows a lack of robustness in some circumstances, but rather with those who make dogmatic assertions of universal robustness in relation to this statistic.

For the sake of exposition, let us now assume that the exaggerated claim of Glass and Stanley (1970) is true. Or even better, let us assume that the _t_-test is perfectly robust to deviations from normality in terms of both type I and type II errors. Even this circumstance would not constitute the convincing argument for exclusive use of the _t_-test that Glass et al. (1972) have taken it to be. These writers state, "The flight to non-parametrics was unnecessary principally because researchers asked 'Are normal theory ANOVA assumptions met?' instead of 'How important are the inevitable violations of normal theory ANOVA assumptions?'" (p. 237). But Glass et al. (1972) miss the important issue, for, as Scheffé (1959, p. 351) has warned, "The question of whether _F_ tests [or in this instance _t_-tests] preserve against non-normal alternatives the power calculated under normal theory should not be confused with that of their efficiency against such alternatives relative to other kinds of tests." The importance of this statement is seen when we realize that the optimal power properties associated with the _t_ statistic are no longer in force when we abandon the requirement that f(x) be normal. In point of fact, there is evidence that large power advantages can be gained by substituting Wilcoxon's test for the _t_-test in non-normal situations. We now turn our attention to some of this evidence.

One of the most common methods used to compare the power of two statistical tests is to compute their asymptotic relative efficiency, abbreviated A.R.E. (sometimes referred to as Pitman efficiency). A.R.E. may be roughly defined as follows:

> Let A and B be two tests based on a and b observations respectively, each test statistic being asymptotically normally distributed (i.e., having a distribution which becomes normal when sample sizes are infinitely large),

> and each testing the same null hypothesis H_0 against the same class of one-sided alternatives $H_a > H_0$, against which both tests are consistent. The A.R.E. of A with respect to B is the limiting value of the ratio b/a as a is allowed to vary in such a way as to give A the same power as B, while simultaneously b approaches infinity and H_a approaches H_0 (Bradley, 1968).

It is interesting to note that the A.R.E. of the Wilcoxon rank-sum test relative to the two-independent means *t*-test is .955 *under normal theory assumptions*. Thus, under this definition of power, the *t*-test shows only a slight power advantage over the Wilcoxon test even when the former test's assumption of population normality is perfectly met. But Glass et al. (1972), as well as many others, contend that the *t*-test is preferable to nonparametric procedures even when f(x) is not normal. A look at some A.R.E.s in this circumstance will be enlightening.

If we assume that f(x) is the logistic distribution, we find that the A.R.E. of the Wilcoxon relative to the *t* is 1.097, indicating a slight power advantage for the Wilcoxon. More interesting is the A.R.E. of 1.5 that is obtained when f(x) is double exponential. Even this rather substantial power advantage of the Wilcoxon statistic pales, however, when we note the A.R.E. of 3 obtained under exponential and gamma distributions (Lehmann 1975; Wetherill, 1960)!

Perhaps the reader has begun to wonder why examples of A.R.E.s that show large (or even moderate) power advantages for the *t*-test have not been included. The answer is simple: There are none. Hodges and Lehmann (1956) have shown that while the A.R.E of the Wilcoxon test relative to the *t*-test can be as large as infinity, it can never be lower than .864. Commenting on this result, Hodges and Lehmann (1956) state:

> To the extent that the above concept of efficiency adequately represents what happens for the sample sizes and alternatives arising in practice, this result shows that use of the Wilcoxon test instead of the Student's _t_-test can never entail a serious loss of efficiency for testing against shift. (On the other hand, it is obvious...that the Wilcoxon test may be infinitely more efficient than the _t_-test.) (p. 356)

At this point we can begin to appreciate the admonition by Scheffé (1959) that was quoted above.

For all of their usefulness as indices of the relative power of two tests, A.R.E.s suffer from at least two major shortcomings. In order to gain their general applicability under a given function, unrealistic assumptior concerning sample size and the condition of the alternatives must be made. As Bradley (1968) has pointed out, "No experimenter takes infinitely large samples, and virtually no one is interested in power to reject hypotheses that differ only infinitesimally from the null hypothesis" (p.58). We will therefore wish to examine the situation in which sample sizes are finite and differences between null and alternative conditions are not restricted in the manner used to compute A.R.E.s. Unfortunately, the evidence in this realm is both limited and specific to the experimental conditions--i.e., relative size of θ, magnitude and location of α, sample sizes as well as other factors. But limited evidence is preferred to unwarranted speculation.

Lehmann (1975) (in tables taken from Dixon (1954) and Hodges and Lehmann (1956)) has shown that when the Wilcoxon statistic is computed on samples of size $n_1 = n_2 = 5$ which have been drawn from a normally distributed population, $\alpha = 4/126$ and θ is allowed to vary in such a way as to allow power to range from .072 to .998; the efficiency of the Wilcoxon test relative to the _t_-test is .968, .978, .961, .956, .960, .960, .964 and .976 for selected values of θ. Thus, in this non-asymptotic example, there is still little difference between the efficiencies of the two statistics.

Neave and Granger (1968) compared the power of the *t* and Wilcoxon statistics by drawing samples of size $n_1 = n_2 = 20$ and $n_1 = 20$, $n_2 = 40$ from approximately normal distributions that differed only in their values of μ, computing the two statistics of interest and recording the proportion of times the null hypothesis was rejected by each test. The difference between these proportions was about .01 in favor of the *t*-test.

Other studies similar to the last two are available, but they merely repeat the result obtained there. The fact is that when samples are finite and drawn from normally distributed populations, there is very little difference between the powers of the two tests. But what happens when f(x) is not normal and samples are finite?

Boneau (1962) found little difference between the powers of the two tests being considered when samples were drawn from rectangular and exponential distributions. The slight advantages that did develop were most frequently, though certainly not always, in favor of the *t*-test. Toothaker (1972) in a study similar to Boneau's obtained essentially the same results. Sample sizes employed in these two studies were generally ≤ 5 though Boneau did use larger sample sizes in a few cases. It should be noted that Blair, Higgins and Smitley (1978) have criticized the use of very small samples in studies of this type since (1) educational research usually involves much larger sample sizes, and (2) results obtained from very small samples often do not carry over to more moderate sample sizes.

In a second part of their study, Neave and Granger (1968) compared the power of the two tests of interest under a form of non-normality that is created by the super-position of two normal distributions. Sample sizes were the same as those mentioned previously. These authors concluded (p. 509) that the Wilcoxon test is "much superior" to the *t*-test when samples are drawn from this particular form of non-normal population. The difference between the proportion of false null hypotheses rejected by the two tests waa as high as .12. Whether or not a power advantage of .12 indicates that one test is "muc superior" to another is largely a matter of individual perspective. However, this figure is impressive when it is compared to the power advantages obtaine by the *t*-test in the studies cited above. (Preliminary results from research in progress by this author and others indicates that when moderate-sized sampl are taken from a rectangular population, the *t*-test may show a power advantage slightly higher than that reported by Boneau (1962) and Toothaker (1972).)

Blair et al. (1978) drew samples of sizes (3,9), (6,6), (9,27), (18,18), (27,81) and (54,54) from an exponential distribution in order to compare the power of the *t* and Wilcoxon tests under this function. Results showed *very large* power advantages in favor of the Wilcoxon statistic. Differences in proportions of null hypotheses rejected were as high as .43, with values betwee .3 and .4 being quite common. (Preliminary results of research in progress by this author and others show substantial power advantages for the Wilcoxon test under various other distributions with the double exponential and truncate normal being two notable examples.)

Although the evidence is too scant to allow firm conclusions, it appears that the results obtained under asymptotic theory are reflected in the finite sample size situation. That is, there is no evidence that the *t*-test ever shows more than a modest power advantage over the Wilcoxon statistic but there *is* evidence that the Wilcoxon can show *very large* power advantages over the *t*-test.

With this discussion in mind, we can see that the contention of Glass et al. (1972) that the most important issue is whether or not the type I and type II error rates of the *t*-test are affected by non-normality, is false.

Although this issue is certainly of great interest, the more important issue deals with whether or not an equally (or more) valid statistic exists that tends to show large power advantages over the t when we relax the requirement that f(x) be normal. Available evidence answers this question in the affirmative and points to the Wilcoxon test as one such statistic.

At this point the rather naive objection might be raised that educational data are rarely sufficiently non-normal to warrant concern. Perhaps the most effective means of dealing with such a notion on the part of an educational researcher is to suggest that he/she routinely construct relative frequency histograms of the data used in statistical analyses. This time-honored but often neglected practice usually paints pictures of distributions unimagined by the researcher who thinks of data in terms of the normal curve. Figures 1-3 are relative frequency histograms of data gathered in connection with an educational study. Distributions could have been presented that are more radically non-normal (in terms of skew for example) than the three exhibited here, but these are of particular interest because they are examples of shapes that tend, in the experience of this researcher, to reoccur in educational data. Ceiling effects, floor effects, presence of large minority groups, special scoring conventions, as well as interactions between these phenomena are only some of the factors that give rise to bizarre shapes in educational data. Although educational data are often roughly normal in appearance, they are also often "heavy-tailed," "light-tailed," "mixed normal," "truncated normal," "L shaped" and "J shaped" in appearance. All of these forms have implications for the validity of the t-test and/or the relative power of the two statistics under discussion.

Summarizing, the major points made thus far are as follows: (1) Statements concerning the robustness of the two independent means t-test are at times highly exaggerated and rarely sufficiently qualified so as to conform with known facts; (2) Although the issue of the robustness of the *t*-test to population non-normality is important, the more important question deals wit whether or not an equally or more valid test exists that tends to demonstrat power superior to that of the *t* statistic when f(x) is not required to be normal; (3) There appears to be no evidence, either in the form of statistica theory or empirical demonstration, to indicate that the *t*-test ever enjoys more than a modest power advantage over the Wilcoxon statistic; (4) There *is* evidence, both in the form of statistical theory and empirical demonstration, to indicate that the Wilcoxon statistic can enjoy *large* power advantages over the *t*-test; and (5) Educational data are often distributed in a radically non-normal manner, thus making the topic under discussion an important one for researchers in education.

Two last points are in order. First, the reader should not be left with the impression that a mirror-image of the position taken by Boneau (1960, 1962) Glass et al. (1972) and many others is being taken here. That is, it is not being suggested that the Wilcoxon test be used exclusively. When it is known that the population form is one that favors the *t*-test or that a contaminated shift is likely or that a large number of tied observations are present, the *t*-test is probably the more appropriate of the two statistics. For general purposes, however, there is little to lose and much to gain by using the Wilcoxon test.

Finally, researchers who apply statistical techniques in the course of educational inquiries are not a "herd" in danger of being frightened into a "stampede" to non-parametric statistics as Glass et al. (1972) have characterized them. They are, however, rational professionals who, when provided with unexaggerated facts and a clear understanding of the issues involved, will choose the most appropriate statistical technique for a given research problem. This is true whether the most appropriate statistical technique happens to be parametric or nonparametric in nature.

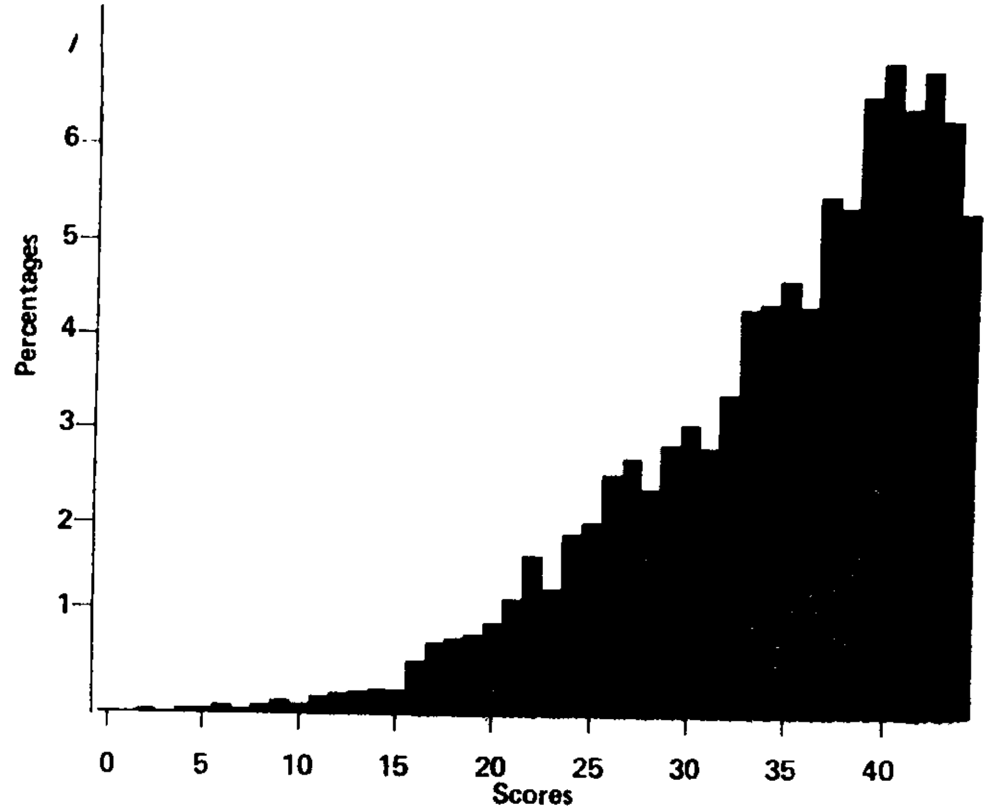

Figure 3. Scores of 7782 fourth grade students on the Comprehensive Tests of Basic Skills – Language-Spelling

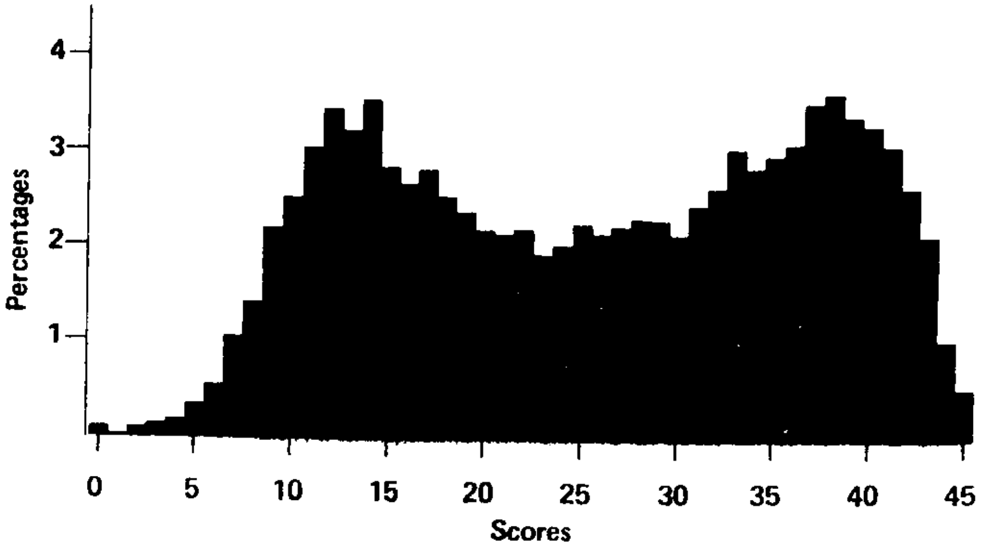

Figure 2. Scores of 8363 third grade students on the Comprehensive Tests of Basic Skills – Reading Comprehension