Journal of Educational and Behavioral Statistics
Summer 2000, Vol. 25, No. 2, pp. 101–132

A Critique and Improvement of the *CL* Common Language Effect Size Statistics of McGraw and Wong

András Vargha
Department of Experimental Psychology, ELTE, Hungary
Harold D. Delaney
Department of Psychology, UNM, U.S.A.

Keywords: *group comparison, effect size measures, measure of stochastic superiority, stochastic equality, stochastic homogeneity, Mann-Whitney-Wilcoxon test, Kruskal-Wallis test, Sign-test, Friedman test*

McGraw and Wong (1992) described an appealing index of effect size, called CL, *which measures the difference between two populations in terms of the probability that a score sampled at random from the first population will be greater than a score sampled at random from the second. McGraw and Wong introduced this "common language effect size statistic" for normal distributions and then proposed an approximate estimation for any continuous distribution. In addition, they generalized* CL *to the* n*-group case, the correlated samples case, and the discrete values case.*

In the current paper a different generalization of CL, *called the* A *measure of stochastic superiority, is proposed, which may be directly applied for any discrete or continuous variable that is at least ordinally scaled. Exact methods for point and interval estimation as well as the significance tests of the* A = .5 *hypothesis are provided. New generalizations of* CL *are provided for the multi-group and correlated samples cases.*

Suppose that an experimenter wishes to assess the difference between two populations with respect to a variable X. If X can be measured on an interval scale, and if the expected value can appropriately characterize the level of the variable in the two populations, then one can measure this difference with the $\mu_1 - \mu_2$ difference of the corresponding expected values. If one is interested in a standardized difference, where the variability is also taken into account, one can use the expression $(\mu_1 - \mu_2)/\sigma$. Here σ is either the common or the pooled *SD*

Much of the work reported herein resulted from the collaborative efforts of Drs. Vargha and Delaney while Vargha was at the University of New Mexico under the support of a Fulbright grant from the U.S. government, and while Vargha was a Széchenyi Professor Scholar and supported by Hungarian OTKA, grant No.: T018353. This work was also supported by the Research Support Scheme of the Open Society Support Foundation, grant No.: 584/1998.

of the two populations (see Cohen, 1977, p. 20; Maxwell & Delaney, 2000; Wilcox, 1996, p. 156-159). Arguing that these and other effect size statistics (like those due to Friedman, 1968; Levy, 1967; Rosenthal & Rubin, 1982) are not well-suited for "communicating effect size to audiences untutored in statistics," McGraw and Wong (1992) proposed a "common language statistic," called *CL*, which converts an effect size into a probability. For continuous distributions *CL* "is the probability that a score sampled at random from one distribution will be greater than a score sampled from some other distribution" (1992, p. 361). Using the conventional notation of probability theory,

$$CL = P(X_1 > X_2), \tag{1}$$

where X_1 is a randomly selected score from the first population (distribution) and X_2 is a randomly selected X-score from the second population (distribution).

McGraw and Wong (1992) in their paper: (1) gave a formula for the point estimation of *CL*, provided that X is normally distributed and $\sigma_1 = \sigma_2$; (2) suggested that *CL* also be used in the case of nonnormal distributions and provided some evidence that *CL* is robust to violations of normality and variance homogeneity; (3) generalized *CL* to the multi-group case, correlated samples case, and discrete case.

In the current paper we propose a different generalization of *CL*, called the *measure of stochastic superiority*, and denoted by A, as follows:

$$A_{12} = P(X_1 > X_2) + .5P(X_1 = X_2). \tag{2}$$

Here the indexes of A refer to the two populations to be compared. A applies for any, not necessarily continuous, distribution that is at least ordinally scaled, and clearly A is identical with *CL* in the continuous case. If X is discrete, A can be interpreted as an estimate of the value of *CL* that would be obtained if the distribution of X were continuous.

One can very easily verify that

$$A_{12} = 1 - A_{21} \tag{3}$$

always holds. If X has the same distribution in the two populations, then from (3) one gets

$$A_{12} = A_{21} = .5, \tag{4}$$

but the reverse implication is not necessarily true. However, if identity (4) holds, then we can conclude that neither population generally has larger X-values than the other. For this reason in this special case we say that the two populations are *stochastically equal* to each other (with respect to X).

From (2) and (4) it is readily seen that stochastic equality is equivalent to the following identity as well (also without the restriction of continuity):

$$P(X_1 > X_2) = P(X_1 < X_2). \quad (5)$$

In the first section of the paper it will be shown that A is a simple but fully adequate generalization of CL that applies equally well to discrete and to continuous distributions. We will also explain here the interpretation for A.

In section 2 we will introduce an unbiased point estimate of A, (1) that is just as easy to compute as the one described by McGraw and Wong for CL, (2) that does not rely on the assumptions of normality and variance homogeneity, and (3) that can be applied for any variable that is at least ordinally scaled.

In section 3 we will provide several test procedures for testing the $A = .5$ hypothesis, i.e., stochastic equality, an issue completely ignored in the paper of McGraw and Wong.

In section 4 we will show some methods for constructing confidence intervals for A.

In section 5 we will show how inadequate the generalization of CL to the multi-group case proposed by McGraw and Wong is, and offer a correct and sound alternative along with the corresponding significance tests.

In the last section we will provide similar improvements with regard to the correlated samples case.

With this paper our primary attempt was not the criticism of McGraw and Wong (1992). We greatly appreciate their effort in introducing the $CL = P(X > Y)$ effect size statistic, which may be more easily understood by many social science researchers than the well known parametric effect size measures. But CL is appropriate for measuring the difference between two distributions only in the continuous case. Realizing that the discrete distributions play a major role in psychology and other behavioral and social sciences (see, e.g., Micceri, 1989), we propose the A measure of stochastic superiority as a direct generalization of CL that is meaningful for any ordinally scaled variable.

1. The Rationale and Interpretation for *A*

In the process of generalization of CL for any, not necessarily continuous distribution that is at least ordinally scaled, our starting point is the formula of stochastic equality defined by (5). Let us denote the left-hand side probability of identity (5) by p_+, and the right-hand side probability by p_-. If identity (5) does not hold because $p_+ > p_-$, $X_1 > X_2$ will occur more often than $X_1 < X_2$, and for this reason in this case we will say that population 1 is *stochastically larger* than population 2 with respect to variable X. Analogously, if instead of (5), rather $p_+ < p_-$ is true, in which case $X_1 < X_2$ will occur more often than $X_1 > X_2$, we will say that population 1 is *stochastically smaller* than population 2 with respect to X.

The concept of "stochastically smaller" or "larger" is not new in the statistical literature (see, e.g., Mann & Whitney, 1947; Kendall & Stuart, 1973, p. 513; Lehmann, 1975, p. 66; Randles & Wolfe, 1979, p. 130; Wilcox, 1990). However, most authors use a stronger form of the *stochastically smaller* relation than that

defined by (4) or (5). According to their definition, X is stochastically smaller than Y if for the respective F_X and F_Y cumulative distribution functions

$$F_X(c) > F_Y(c) \text{ for all } c \tag{6}$$

holds. It can be seen that (6) always implies our weaker form, while the opposite is not necessarily true (see Randles & Wolfe, 1979, p. 132).

For continuous distributions $p_+ = 1 - p_-$, and thus in this case stochastic equality is readily seen to be equivalent to the

$$p_+ = .5 \tag{7}$$

identity. Also in the continuous case $p_+ > .5$ clearly implies that population 1 is stochastically larger than population 2, and $p_+ < .5$ implies that population 1 is stochastically smaller than population 2. These facts show that in the continuous case the $CL = p_+$ measure can clearly indicate in itself the stochastic smaller, equal, and larger relations of populations 1 and 2 with respect to X.

However, if X is discrete, a single p+ value cannot generally inform us about which of populations 1 and 2 is stochastically larger than the other, unless we know also the $p_e = P(X_1 = X_2)$ probability. For example, $p_+ = .3$ can occur with population 1 being stochastically larger than 2 ($p_+ = .3$, $p_e = .6$, $p_- = .1$), or equal to 2 ($p_+ = .3$, $p_e = .4$, $p_- = .3$), or smaller than 2 ($p_+ = .3$, $p_e = .1$, $p_- = .6$). Accordingly, in the case of discrete variables both p_+ and p_- need to be known in determining the stochastic order (relation) of populations 1 and 2.

Obviously, the simplest measure that takes into account both p_+ and p_- is the

$$\delta = p_+ - p_- \tag{8}$$

difference (see Cliff, 1993, p. 494), which may be called the *stochastic difference* of populations 1 and 2 with respect to X. Values of δ can run from -1 (nonoverlapping distributions with $p_+ = 0$ and $p_- = 1$) through 0 (stochastic equality of populations 1 and 2 with $p_+ = p_-$) to 1 (nonoverlapping distributions with $p_+ = 1$ and $p_- = 0$). If we apply a simple linear transformation on δ to attain transformed values that vary in the interval [0, 1] instead of [−1, 1], just like CL in the continuous case where the .5 middle point corresponds to the stochastic equality of populations 1 and 2, one gets:

$$\delta' = (\delta + 1)/2 = (p_+ - p_- + 1)/2 = [p_+ - (1 - p_+ - p_e) + 1]/2 = \\ = (2p_+ + p_e)/2 = p_+ + .5p_e = A_{12},$$

that is

$$A_{12} = (\delta + 1)/2. \tag{9}$$

From (9) it is now obvious that the A_{12} measure of stochastic superiority is the unique linear transformation of the $\delta = p_+ - p_-$ difference that yields the CL common language effect size statistic in the continuous case.

From (9) it follows directly that

$$\delta = 2A_{12} - 1, \tag{10}$$

which helps us interpret the A_{12} values. For example if we have $A_{12} = .65$, then $\delta = 2(.65) - 1 = .3$. This means that in this case $p_+ - p_- = .3$, i.e., the probability that a random X_1 score from population 1 will be greater than a random X_2 score from population 2 is larger than the opposite $P(X_1 < X_2)$ probability by just .3.

Another way of assessing the magnitude of particular A_{12} values is as follows. Suppose first that the variable X is normally distributed in both populations to be compared with expected values μ_1 and μ_2 and standard deviations $\sigma_1 = \sigma_2 = \sigma$. Then let us introduce the $\Delta = (\mu_1 - \mu_2)/\sigma$ notation. According to Cohen one can say that the Δ effect size (i.e., the standardized difference between populations 1 and 2) is small if $\Delta = .2$, medium if $\Delta = .5$, and large if $\Delta = .8$ (see Cohen, 1977, p. 26 or Wilcox, 1996, p. 157). Following the instructions of McGraw and Wong (1992, Figure 1) one can easily calculate in this continuous case the $CL = p_+ = A_{12}$ as well as the $\delta = p_+ - p_- = 2p_+ - 1$ values corresponding to the above mentioned three different levels of effect size, which yields the values listed in Table 1. Even if X is discrete, these tabled values may help us in judging whether a particular value of A_{12} reflects a slight, a moderate, or a pronounced superiority of population 1 compared to population 2 with respect to variable X.

The Generalization of CL to the Discrete Case as Suggested by McGraw and Wong

For dichotomous X variables (with possible values say a and b) McGraw and Wong (1992, p. 363) suggested the following generalization of CL (denoted temporarily by CLg):

$$CLg = P(X_1 = a, X_2 \neq a) = P(X_1 = a)P(X_2 = b).$$

If X is polichotomous, with possible values of $a_1, \ldots a_k$, then McGraw and Wong suggested the same approach, "which is to multiply the separate event probabilities" (McGraw and Wong, 1992, p. 363). This certainly implies that for any possible a_j value of X ($j = 1, \ldots, k$)

$$CLg = CLg(a_j) = P(X_1 = a_j, X_2 \neq a_j) = \Sigma_{j \neq i} P(X_1 = a_j)P(X_2 = a_i).$$

We have the following reservations with this issue. First, this definition is not a proper generalization of CL, because CL is clearly not a special case of CLg (CLg does not reduce to CL when $k = 2$). In addition, CLg can hardly characterize the dominance relation of two populations since this definition of CLg does not require even the ordinality of variable X. Finally, it is very inconvenient that for a k-valued discrete variable the formula does not yield a single measure but rather a series of $CLg(a_j)$ values ($i = 1, \ldots k$), i.e., a k-tuple.

TABLE 1
Guidelines for interpreting A_{12} and δ: corresponding values for normal distributions

	Effect size		
	Small	Medium	Large
$\Delta = (\mu_1 - \mu_2)/\sigma$:	.2	.5	.8
$A_{12} = p_+ + .5p_=$:	.56	.64	.71
$\delta = p_+ - p_-$:	.11	.28	.43

Note: A_{12} and δ values are rounded to two decimals.

To demonstrate how obscure the interpretation of this generalization is, let us have a 3-valued ordinal variable X with possible values of 1, 2, and 3. Suppose that the distribution of X in population 1 is such that these values occur with probabilities .1, .6, and .3, respectively, and that the corresponding probabilities indicating its independent distribution in population 2 are .3, .6, and .1, respectively. Population 1 is clearly superior to 2, because

$$p_+ = P(X_1 > X_2) = (.6)(.3) + (.3)(.9) = .18 + .27 = .45, \text{ and}$$
$$p_- = P(X_1 < X_2) = (.1)(.7) + (.6)(.1) = .07 + .06 = .13,$$

which means that X_1 dominates X_2 more often than X_2 does X_1. However, if we use the definitions of McGraw and Wong for *CL* and *CLg* we have:

$$CL = p_+ = .45,$$
$$CLg(1) = P(X_1 = 1, X_2 \neq 1) = (.1)(.7) = .07,$$
$$CLg(2) = P(X_1 = 2, X_2 \neq 2) = (.6)(.4) = .24,$$

and

$$CLg(3) = P(X_1 = 3, X_2 \neq 3) = (.3)(.9) = .27.$$

Which of these values can indicate alone or jointly with some others that variable X generally has larger values in population 1 than in population 2? However, in this case

$$\delta = p_+ - p_- = .45 - .13 = .32 > 0$$

and

$$A_{12} = (\delta + 1)/2 = 1.32/2 = .66 > .5$$

(see identities (8) and (9)), both measure showing clearly a considerable superiority of population 1 over population 2 (see also Table 1).

2. The Point Estimation of *A*

Suppose a variable of at least ordinal scale is defined for two populations. Assume this variable is denoted by X in population 1 and by Y in population 2. Then the stochastic superiority of population 1 over population 2 would be defined as

$$A_{12} = P(X > Y) + .5P(X = Y). \tag{11}$$

Now assume $X_1, X_2, \ldots, X_m$ is a random sample of size m of values of X drawn from population 1 and $Y_1, Y_2, \ldots, Y_n$ is a random sample of size n of values of Y, drawn independently from the first sample. Since the $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$ observations are all independent from each other, we can estimate A_{12} from these sample as follows:

$$\hat{A}_{12} = \#(X_i > Y_j)/nm + .5\#(X_i = Y_j)/nm. \tag{12}$$

Here $\#(X_i > Y_j)$ denotes the occurrences in the samples of the event $X_i > Y_j$, and $\#(X_i = Y_j)$ denotes the number of occurrences of the event $X_i = Y_j$ (here $i = 1, \ldots, m$, and $j = 1, \ldots, n$). Due to the known relationship between the relative frequency and the probability of an event, $\hat{A}$, as defined by (12), can be considered an unbiased estimate of A, defined by (11).

Based on (12), $\hat{A}$ can be determined very easily. However, if we perform a rank transformation on the original scores, as is done in the Mann-Whitney-Wilcoxon test (see, e.g., Wilcox, 1996, p. 365), and realize that the rank sum of sample 1 can be expressed as follows:

$$R_1 = m(m + 1)/2 + \#(X_i > Y_j) + .5\#(X_i = Y_j), \tag{13}$$

(see, e.g., Maritz, 1995, p. 91), from (12) and (13) we can derive the following simpler formula for $\hat{A}$:

$$\hat{A}_{12} = (R_1/m - (m + 1)/2)/n. \tag{14}$$

Because the Mann-Whitney-Wilcoxon test, along with the R_1, R_2 rank sum values, is available in most statistical program packages (BMDP, SPSS, SAS, etc.), the computation of $\hat{A}$, based on (14), requires only a couple of elementary computations.

Compared to the estimation of *CL* proposed by McGraw and Wong (1992), the present method does not impose any restriction on X beyond ordinality (neither normality nor continuity is assumed), and, specifically, our method does not restrict the form of the distribution of X, does not need the assumption of variance homogeneity, and is always accurate (unbiased).

Example 1

Suppose we have two independent samples of size $m = 15$ and $n = 20$ taken at random from populations P and Q respectively. Suppose furthermore that,

ranking the two data samples together (as in the Mann-Whitney-Wilcoxon test), the rank sum of sample 1 is $R_1 = 330$. Then using formula (14) we get the following unbiased estimate of the A measure of stochastic superiority:

$$\hat{A}_{PQ} = (330/15 - 16/2)/20 = (22 - 8)/20 = .70.$$

This means that in this case the chance that a random score X from population P will be greater than an independently drawn random score Y from population Q plus one half times the change that X and Y will be equal is approximately .70. Also we have $p_* - p_- = 2(.70) - 1 = .40$ (see identity (10)), that is, in this case $X > Y$ occurs much more frequently than $X < Y$. Based on Table 1 we can conclude that this level of $\hat{A}_{PQ}$ reflects a substantial superiority of population P over population Q.

3. Testing the $A = .5$ Hypothesis

The $A = .5$ hypothesis, that is the hypothesis of stochastic equality (see 4)), can be tested by using any of several methods.

The oldest test that is closely related to stochastic equality is the Mann-Whitney-Wilcoxon rank sum test (abbreviated as *MWW* test in the rest of our paper). It is well known that if X is a continuous variable, the null hypothesis that X has the same distribution in populations 1 and 2 can validly be tested with this method (Mann & Whitney, 1947; Noether, 1967, p. 31; Gibbons & Chakraborti, 1992, p. 221; Wilcox, 1996, p. 365). For continuous distributions the *MWW* test has been shown to be consistent with respect to the

$$CL = A_{12} = p_* = P(X_1 > X_2) \neq .5$$

alternative hypothesis (Kendall & Stuart, 1973, p. 513; Miller, 1986, p. 52), that is stochastic inequality. For any, not necessarily continuous, distribution the *MWW* test has been shown to be consistent for an alternative hypothesis if and only if it implies

$$p_* - p_- = P(X_1 > X_2) - P(X_1 < X_2) \neq 0$$

(Noether, 1967, p. 35-36), that is, if populations 1 and 2 *are not* stochastically equal with respect to variable X (see identity (5)). Accordingly, we can conclude that for any discrete or continuous variable X the *MWW* test is consistent against an alternative hypothesis if and only if the two populations to be compared are stochastically unequal with respect to X. This implies that if population 1 and 2 are stochastically unequal, then at any fixed α level of significance the probability of the *MWW* test being significant at the α level tends to 1 as the m, n sample size values tend to infinity. By contrast, if populations 1 and 2 are stochastically equal, then the probability of the *MWW* test being significant at the α level will not tend to 1, however large the two samples be.

As a consequence we can say that although the null hypothesis of the *MWW* test is the equality of the whole distributions, the test is especially sensitive to one aspect of difference, namely, stochastic inequality.

In order to reduce the generality of this null hypothesis, several authors assume that the two distributions to be compared are identical except possibly for their locations. The model with this restriction is known as the additivity or shift model (see Hájek, 1969, p. 41; Lehmann, 1975, p. 66; Hettmansperger, 1984, p. 132-133; Maritz, 1995, p. 79-80). With this stringent assumption the equality of the two distributions will obviously be equivalent to the equality of the two populations to be compared, as well as the equality of the expected values (if they exist), and the equality of the medians (only if X is continuous).

The question arises to what extent can this stringent assumption of additivity be weakened with the *MWW* test being still valid? Several simulation studies have indicated that, similarly to the two-sample t test, the *MWW* test is not robust to unequal population variances, especially in the unequal sample size case (see for example Zimmerman & Zumbo, 1992, 1993a).

The failure of the *MWW* test under variance heterogeneity motivates a search for robust methods to test legitimately the null hypothesis of stochastic equality without any restrictive side condition upon the distribution of X in populations 1 and 2. We found three methods in the statistical literature that may offer an answer to this problem.

(a) For testing the equality of medians of two continuous distributions, Fligner and Policello (1981) published a modified *MWW* test procedure (denoted in the following as the *FP* test), which did not assume the equality of population variances. Using the notations of section 2 the computation of the *FP* test is as follows. For each score, compute a *placement score* (which we will denote as V_i, $i = 1, \ldots, m$, for the X-sample and W_j, $j = 1, \ldots, n$ for the Y-sample), where V_i equals the number of Y-scores less than X_i ($i = 1, \ldots, m$), and W_j equals the number of X-scores less than Y_j ($j = 1, \ldots, n$). Then the test statistic of the *FP* test is

$$Z_{FP} = d/s_d, \tag{15}$$

where the d numerator, a point estimate of $\delta = p_+ - p_-$, is defined as follows:

$$d = (\Sigma_i V_i - \Sigma_j W_j)/(mn), \tag{16}$$

and the s_d denominator, an estimate of the SD of d, using the notation $v = \Sigma_i V_i/m$, $w = \Sigma_j W_j/n$, $SS_V = \Sigma_i (V_i - v)^2$, $SS_W = \Sigma_j (W_j - w)^2$, is defined as follows:

$$s_d = 2(SS_V + SS_W + vw)^{1/2}/(mn) \tag{17}$$

(see Wilcox, 1996, p. 369). In a simulation study Zumbo and Coulombe (1997) have shown that the *FP* test is generally not robust to the assumption of

symmetric distributions if the equality of medians is tested. However, Fligner and Policello (1981, p. 164) assert that if instead of testing the equality of medians "we were interested in testing H_0: $\int GdF = .5$" (which is equivalent to $P(X_1 > X_2) = .5$ and stochastic equality; see Randles & Wolfe, 1979, p. 132), we could use the *FP* test without the symmetry assumption. Wilcox introduces the *FP* test also as a method capable to test H_0: $P(X_1 > X_2) = .5$ and mentions that it is an improvement of the *MWW* test that eliminates the assumption of equal variances (1996, p. 369-371).

Cliff (1993) suggests that the *FP* test be used for testing the null hypothesis

$$H_0: \delta = P(X_1 > X_2) - P(X_1 < X_2) = 0$$

that is stochastic equality (see (5)), without assuming identical distributions (which implies among other things the equality of variances). The reason for this is that in the Z_{FP} test the d statistic is an unbiased estimate of the $\delta = p_+ - p_-$ stochastic difference and s_d, given by formula (17), is a consistent estimator of σ_d, the standard deviation of d (see Cliff, 1993, p. 499). We note that $\delta = 0$ is equivalent to stochastic equality without the restrictive assumption of continuity (see (5)).

(b) Cliff suggested an alternative to Z_{FP} by replacing s_d with S_d, a different estimator of σ_d, which is defined by

$$(S_d)^2 = \frac{n^2\Sigma_i(d_i - d)^2 + m^2\Sigma_j(d_j - d)^2 + \Sigma_i\Sigma_j(d_{ij} - d)^2}{mn(m-1)(n-1)} \tag{18}$$

(see Cliff, 1993, identity (9)). Here $d_{ij} = \text{sign}(X_i - Y_j)$, $d_i = \Sigma_j d_{ij}/n$, $d_j = \Sigma_i d_{ij}/m$, and d, the average of all d_{ij} values, is the same as in (16) (sign(c) of any number c is defined to be -1, 0, or 1, if c is negative, zero, or positive, respectively).

Both these tests rely on the asymptotic standard normality of their test statistics. However, Fligner and Policello (1981, Table 1) report also exact significance levels when both sample sizes are between 3 and 12 (see also Wilcox, 1996, Table 14 of Appendix B).

(c) A third test procedure to be proposed for testing the null hypothesis of stochastic equality is Welch's t test (see Wilcox, 1996, p. 133), performed on the rank scores of the *MWW* test (Zimmerman & Zumbo, 1992, 1993a, 1993b). The rationale for the applicability of this test is as follows.

If we take the expected value of both sides of (14) we get

$$A_{12} = [E(R_1/m) - (m+1)/2]/n. \tag{19}$$

In (19) $E(R_1/m)$ is the expected value of the rank mean of sample 1, and, as the X-scores are identically distributed, it equals also the common expected value of the rank scores in sample 1. According to (19) the $A_{12} = .5$ equality is equivalent to the identity

$$E(R_1/m) = (n + m + 1)/2. \tag{20}$$

Similarly, the $A_{21} = .5$ equality is equivalent to the identity

$$E(R_2/n) = (n + m + 1)/2. \tag{21}$$

From (19), (20), and (21) it follows that the defining identity of stochastic equality:

$$H_0\text{: } A_{12} = A_{21} = .5 \tag{22}$$

is equivalent to the following identity as well:

$$H_0'\text{: } E(R_1/m) = E(R_2/n). \tag{23}$$

Since this latter hypothesis states that the rank means, and consequently the rank scores, have identical expected values in the two samples, stochastic equality can be tested by a two-sample mean comparison method. The most commonly used test for this purpose is Student's two-sample *t* test, which has the following three assumptions: (1) independence of all individual observations from each other (this implies also the independence of the two samples), (2) normality of the parent distributions, and (3) variance homogeneity. The first assumption holds asymptotically for the rank scores, since there is only one single constraint that makes them slightly correlate (negatively) with each other: they always sum to $N(N+1)/2$. The second and third assumption may frequently be violated in empirical studies (see, e.g., Micceri, 1989; Wilcox, 1996, p. 135), and this may invalidate the *t* test (Wilcox, 1996, p. 135), as well as the *t* test performed on the rank scores (the rank *t* test). Because one may not have available a good test for checking variance homogeneity, and because a robust alternative, namely Welch's *t* test, generally improves upon Student's *t* test under the violation of normality and variance homogeneity (Wilcox, 1996, p. 135), it seems to be a reasonable choice to perform this robust alternative to *t* on the rank scores (rank Welch test). But taking into account also the fact that in certain situations Welch's method may be considerably inflated, it is important to identify the circumstances under which the rank Welch test is an appropriate test of stochastic equality.

A Monte Carlo Study

A Monte Carlo study provided empirical information about the behavior of the three procedures described above. With this study we wanted to obtain some evidence that these methods are acceptably robust to variance heterogeneity when testing the null hypothesis of stochastic equality. This simulation study seemed to be especially important, because nobody has published extensive validity results concerning statistical tests of stochastic equality for asymmetric distributions, where the equality of location parameters (means or medians) does not necessarily imply stochastic equality, and vice versa. Nevertheless, some evidence has been accumulated from simulation studies with the *FP* and the

rank Welch tests for testing the null hypothesis of means or medians involving symmetric distributions.

Zimmerman and Zumbo (1993a) showed, for example, that for normal distributions "the Welch t' test protected against changes resulting from unequal variances in combination with unequal sample sizes, not only when it was performed on the initial scores, but also when it was performed on the ranks of the scores" (1993a, p. 531). Similar results, with slightly inflated Type I error rates, were obtained also for some other symmetric distributions involving the mixed normal, Cauchy, Laplace, uniform, and mixed uniform distributions (Zimmerman & Zumbo, 1992, 1993a). Other results of this investigation involving such asymmetric distributions as the exponential, half-normal, and lognormal, cannot be directly interpreted in terms of stochastic equality. The reason for this is that for asymmetric distributions a transformation that will equate the means coupled with different variances will not necessarily equate stochastically the two distributions, just as it will not equate the medians (see the footnote on p. 143 of Zumbo & Coulombe, 1997).

Fligner and Policello (1981) reported some simulation results concerning the robustness of the *FP* test against variance heterogeneity for several symmetric continuous distributions (where again the tested equality of medians was equivalent to stochastic equality). They found that their test "maintained its nominal level well for all situations considered" (1981, p. 167). Zumbo and Coulombe (1997) also carried out a simulation study with the *FP* test for testing the equality of two medians with small samples ($n, m \leq 12$), with samples drawn from either a symmetric (normal) or heavily skewed (ex-Gaussian) parent distribution type. An ex-Gaussian distribution is defined as the sum of a normally distributed (with parameters μ and σ) and an independent and exponentially distributed random variable (with parameter t). These parameters specify the shape of the ex-Gaussian. Miller (1988) listed several combinations of parameter values representative of those found in empirical studies of reaction time. For their simulation study Zumbo and Coulombe (1997) selected Miller's most skewed distribution with parameter values $\mu = 300$, $\sigma = 20$, and $t = 300$. They found that the *FP* test performed very inconsistently for the ex-Gaussian distribution (see Table 2 in Zumbo & Coulombe, 1997), which again may be due to the fact that under variance heterogeneity, equating the medians will generally not achieve stochastic equality, the proper null hypothesis of the *FP* test for asymmetric distributions (see Fligner & Policello, 1981, p. 164). In addition, Zumbo and Coulombe (1997) report that for the normal distribution the *FP* test performs quite conservatively even in cases where Fligner and Policello obtained close to nominal level coverage. For example, for $m = 11$, $n = 10$, and $\alpha = .05$, Fligner and Policello obtained .048 as an estimate for the Type I error rate (see Table 2 of Fligner & Policello, 1981), whereas Zumbo and Coulombe report an estimated value of .030 for the same parameter under the same condition (equal variance, two-sided test; see Table 1 of Zumbo & Coulombe, 1997). With our simulation study we wanted also to find an explanation to this obvious inconsistency.

TABLE 2

Shift values ensuring stochastic euality for the appled lambda distribution types depending on the standard deviations of the two distributions to be compared. The initially standardized X *and* Y *lambda distributions were submitted to the* $X' = \sigma_1 X$, $Y' = \sigma_2 Y + C$ *linear transformations, where* C *is the corresponding tabled shift value*

		(σ_1, σ_2)				
Skew	Kurt	(1, 1)	(1, 2)	(2, 1)	(1, 3)	(3, 1)
0	1.8	0.0000	0.0000	0.0000	0.0000	0.0000
0	3.0	0.0000	0.0000	0.0000	0.0000	0.0000
0	9.0	0.0000	0.0000	0.0000	0.0000	0.0000
2	8.6	0.0000	0.4223	−0.4223	0.7795	−0.7795
2	12.0	0.0000	0.3615	−0.3615	0.6808	−0.6808
2	15.8	0.0000	0.2070	−0.2070	0.3985	−0.3985

For the simulation study we selected four statistics. The rank *t* test, denoted by *rt*, although not supposed to be robust to variance heterogeneity, was considered for comparison purposes. The other three statistics were the above mentioned robust alternatives: the rank Welch test, denoted by *rW*, Fligner and Policello's *FP* test, and Cliff's modified *FP* test, denoted by *FPC*.

Random variates were generated from the generalized lambda family of distributions, which offers a variety of different shapes (Ramberg, Tadikamalla, Dudewicz, & Mykytka, 1979). These distributions are given in standardized form and can be described in terms of skewness ($\alpha_3 = \mu_3/\sigma^3$) and kurtosis ($\alpha_4 = \mu_4/\sigma^4$), where μ_3 and μ_4 are the third and fourth central moments. The generalized lambda family covers a wide range of values of skewness and kurtosis so that for any given value of skewness, several values of kurtosis can be specified (see Table 4 in Ramberg et al., 1979). For the present study two levels of skewness ($\alpha_3 = 0$ and $\alpha_3 = 2$) were applied. For each level of skewness three levels of kurtosis were used (for $\alpha_3 = 0$: $\alpha_4 = 1.8$, 3.0, and 9.0, and for $\alpha_3 = 2$: $\alpha_4 = 8.6$, 12.0, and 15.8). The lowest and highest levels of kurtosis always represent the most extreme levels available in Table 4 of Ramberg et al. (1979). The middle levels correspond to a medium level of kurtosis, which for a symmetric distribution gives a generalized lambda distribution having the first four moments equal to those of the standard normal. The two levels of skewness together with the tree levels of kurtosis for each yielded six different distribution types.

In the two samples (of size *m* and *n*) we considered only identical distribution types, but allowing for differences in variances. The applied variance ratios were either 1:1 or 1:9. The tested null hypothesis was stochastic equality. To achieve this, in the asymmetric distribution, unequal variances case, the second distribution has been shifted by an appropriate constant to the right. The shift constants ensuring stochastic equality were determined empirically by a Turbo Pascal program prior to the simulation process by means of successive iterations until

stochastic equality was fulfilled. The criterion of stochastic equality was specified by the satisfaction of identity (4). A shift value was accepted if the estimated A_{12} value differed from .5 by not more than $\pm$ 0.001 three consecutive times, each time applying 1.6 million randomly generated couples of variable values sampled from the two distributions. The obtained shift values are summarized in Table 2.

The average sample size was either small ($(m+n)/2 = 9$) or moderately large ($(m+n)/2 = 18$). In each case we applied both a balanced ($n = m$) and an unbalanced ($n = 2m$) design. In the unequal samples unequal variances case both the direct and inverse pairing conditions were investigated.

The study was conducted on a Pentium 200 MHz IBM PC compatible computer. The generalized lambda random variates were generated using the method described in Ramberg et al. (1979). In this generation process, Turbo Pascal's Random function was used to obtain pseudo-random uniform deviates. It is a linear congruential random-number generator that has turned out to be one of the most preferable in a recent study (Onghena, 1993), passing successfully ten criterion tests of randomness. For each choice of distributions and sample sizes we employed 100,000 simulation iterations. At each iteration, $N = m + n$ random variates of the desired type were generated. All of the four tests were then performed on the current set of N variates and evaluated in two-tailed form with significance level .05. Test statistics *rt* and *rW* were evaluated according to the usual *t* percentile values (based on the corresponding *df*'s). In the case of *FP* and *FPC*, for small samples ($N = 18$) we used the exact critical points reported by Fligner and Policello (1981), and for moderately large samples ($N = 36$) we used the normal approximation method (see Wilcox, 1996, p. 370). Our computer coding of *rt* and *rW* was checked by applying the Group comparison routine of the Ministat statistical program package (Vargha, 1999) to the numerical example of Example 2 (see below), and *FP* was checked against a computational example found in Wilcox (1996, p. 370–371). Our attempt to check *FPC* against the computational example found in Cliff (1993, p. 500) failed due to a numerical error discovered in Cliff's computation. His Table 1 has an incorrect entry: the "−1" value in the cell in the next to last column three rows from the bottom should be "+1". Finally, the proportion of rejections (out of the 100,000 replications) was determined. The obtained empirical Type I error estimates are reported in Table 3 ($N = 18$) and in Table 4 ($N = 36$). If the true level of the four tests was .05, an entry in Table 3 and Table 4 would have a standard deviation of $.00069 = [(.05)(.95)/100000]^{1/2}$.

The obtained results can be explained as follows:

1. If the two distributions have *equal variances*, *rt* maintains its nominal level over all distribution and sample size settings. The case is similar with *rW*, but it has a slightly increased Type I error level when the sample sizes are small and unequal. *FP* also performs nicely in the small sample settings (where exact critical points were used), but is somewhat inflated when its test statistic is evaluated via the normal distribution table. Under equal variances the worst performing test is *FPC*.

TABLE 3

*Empirical Type I error rates of four rank tests (*rt = *rank* t *test,* rW = *rank Welch test,* FP = *Fligner-Policello test,* FPC = *Cliff's modified* FP *test) for testing the null hypothesis of stochastic equality against two-sided alternatives at 5% nominal level, with an average sample size of 9.* FP *and* FPC *were evaluated with exact tail probabilities (small sample versions)*

	Equal variances		Unequal variances		
				Direct	Inverse
(m, n):	9, 9	6, 12	9, 9	6, 12	12, 6
$(\sigma_1 : \sigma_2)$:	1:1	1:1	1:3	1:3	1:3
Kurtosis					
	Symmetric distributions ($\alpha_3 = 0$)				
$\alpha_4 = 1.8$					
rt:	0.051	0.055	0.077++	0.034−	0.133++
rW:	0.051	0.058	0.066+	0.071+	0.060+
FP:	0.051	0.053	0.060	0.037−	0.063+
FPC:	0.059	0.058	0.063+	0.043	0.063+
$\alpha_4 = 3.0$					
rt:	0.050	0.054	0.069+	0.032−	0.118++
rW:	0.050	0.058	0.065+	0.064+	0.064+
FP:	0.051	0.052	0.057	0.035−	0.067+
FPC:	0.059	0.057	0.063+	0.041	0.066+
$\alpha_4 = 9.0$					
rt:	0.049	0.053	0.066+	0.034−	0.108++
rW:	0.049	0.056	0.063+	0.063+	0.063+
FP:	0.050	0.050	0.056	0.038−	0.065+
FPC:	0.058	0.055	0.062+	0.043	0.064+
	Asymmetric distributions ($\alpha_3 = 2$)				
$\alpha_4 = 8.6$					
rt:	0.050	0.053	0.069+	0.033−	0.117++
rW:	0.050	0.056	0.064+	0.064+	0.065+
FP:	0.050	0.051	0.057	0.036−	0.067+
FPC:	0.058	0.056	0.062+	0.041	0.066+
$\alpha_4 = 12.0$					
rt:	0.050	0.053	0.068+	0.034−	0.115++
rW:	0.050	0.057	0.064+	0.065+	0.063+
FP:	0.050	0.051	0.057	0.037−	0.065+
FPC:	0.059	0.056	0.062+	0.042	0.065+
$\alpha_4 = 15.8$					
rt:	0.050	0.053	0.067+	0.033−	0.111++
rW:	0.050	0.056	0.063+	0.062+	0.065+
FP:	0.050	0.050	0.057	0.036−	0.067+
FPC:	0.059	0.055	0.062+	0.041	0.066+

Note: − Denotes values less than .040

\+ Denotes values exceeding .060

++ Denotes values exceeding .075

TABLE 4

Empirical Type I error rates of four rank tests (rt = rank t test, rW = rank Welch test, FP = Fligner-Policello test, FPC = Cliff's modified FP test) for testing the null hypothesis of stochastic equality against two-sided alternatives at 5% nominal level, with an average sample size of 18. FP and FPC were evaluated with the standard normal approximation (large sample versions)

		Equal variances		Unequal variances		
					Direct	Inverse
	(m, n):	9, 9	6, 12	9, 9	6, 12	12, 6
	$(\sigma_1 : \sigma_2)$:	1:1	1:1	1:3	1:3	1:3
Kurtosis						
		Symmetric distributions ($\alpha_3 = 0$)				
$\alpha_4 = 1.8$						
	rt:	0.050	0.049	0.076++	0.031−	0.116++
	rW:	0.050	0.051	0.072+	0.064+	0.065+
	FP:	0.058	0.064+	0.064+	0.056	0.077++
	FPC:	0.063+	0.069+	0.067+	0.062+	0.076++
$\alpha_4 = 3.0$						
	rt:	0.051	0.049	0.071+	0.030−	0.106++
	rW:	0.051	0.052	0.068+	0.061+	0.063+
	FP:	0.059	0.065+	0.065+	0.055	0.077++
	FPC:	0.064+	0.070+	0.068+	0.061	0.076++
$\alpha_4 = 9.0$						
	rt:	0.051	0.049	0.066+	0.031−	0.099++
	rW:	0.051	0.052	0.064+	0.061+	0.062+
	FP:	0.059	0.065+	0.063+	0.057	0.077++
	FPC:	0.064+	0.070+	0.066+	0.063+	0.076++
		Asymmetric distributions ($\alpha_3 = 2$)				
$\alpha_4 = 8.6$						
	rt:	0.050	0.050	0.070+	0.031−	0.107++
	rW:	0.050	0.053	0.067+	0.062+	0.064+
	FP:	0.059	0.065+	0.064+	0.056	0.078++
	FPC:	0.065+	0.070+	0.067+	0.062+	0.077++
$\alpha_4 = 12.0$						
	rt:	0.051	0.048	0.070+	0.031−	0.104++
	rW:	0.051	0.051	0.067+	0.061+	0.062+
	FP:	0.060+	0.064+	0.064+	0.055	0.077++
	FPC:	0.066+	0.069+	0.067+	0.061+	0.076++
$\alpha_4 = 15.8$						
	rt:	0.050	0.049	0.067+	0.031−	0.099++
	rW:	0.050	0.053	0.065+	0.060+	0.061+
	FP:	0.059	0.065+	0.063+	0.055	0.075+
	FPC:	0.065+	0.070+	0.067+	0.061+	0.074+

Note: − Denotes values less than .040

\+ Denotes values exceeding .060

++ Denotes values exceeding .075

2. In *unequal variances* settings the test performances depend heavily on the correlation of sample size and variance values. Obviously *rt* is the test that is most susceptible to this correlation, which is a well known phenomena (see, e.g., Zimmerman & Zumbo, 1992, 1993a). Quite interestingly, however, *FP* and *FPC* show a similar pattern, though to a lesser degree, showing an increasing trend of Type I error rates when this correlation goes from large negative (inverse pairing of sample sizes and variances) to large positive (direct pairing of sample sizes and variances). At the same time, *rW* has a consistently slight inflation of Type I error. Due to these circumstances, in the direct pairing condition the test that maintains its nominal level with the greatest precision is *FPC* (for small samples) or *FP* (for medium size samples). In the equal n, unequal variances conditions, *FP* is best, and in the inverse pairing condition, *rW*.

3. The large sample versions of *FP* and *FPC* are consistently inflated under all conditions. For this reason we suggest that one calculate and use exact critical values not only for $n, m \leq 12$ but also for $n, m \leq 30$.

4. No perceivable differences in Type I error rates have been found between symmetric and asymmetric distributions. The effect of kurtosis seems also to be minimal. At the lowest level of kurtosis, *rW* shows sometimes higher Type I error values than in other cases and *rt* shows a similar pattern, which is most salient in the inverse pairing condition.

5. Our results concerning *FP* are consistent with those of Fligner and Policello (1981) in the equal variances (and in the equal sample sizes unequal variances) conditions, and consequently do not match those strongly conservative Type I error rates of *FP* under normal parent distribution obtained by Zumbo and Coulombe (1997). The case is just opposite in the direct pairing condition, where our results are consistent with those of Zumbo and Coulombe but contradict those of Fligner and Policello. Finally, in the inverse pairing condition our inflated Type I error results contradict both Fligner and Policello (1981), and Zumbo and Coulombe (1997).

6. For testing the null hypothesis of stochastic equality in the unequal variance conditions the proposed robust rank tests may produce substantially better Type I error rates than does *rt*, the nonparametric analogue of Student's two-sample t test, which in turn seems to be fully adequate in the equal variances condition.

Our primary goal with this small simulation study was to provide some empirical evidence that for testing the null hypothesis of stochastic equality our proposed robust methods may really be able to work without such restrictive side conditions imposed upon the parent distributions as symmetry and variance homogeneity, even if they are sometimes a little inflated. To give full justification to them, however, further investigations are needed with other distribution types (including also discrete ones), sample sizes, and variance ratios, with considerations also given to power comparisons.

4. Interval Estimation of the Measure of Stochastic Superiority

All three robust tests of stochastic equality described in the previous section can be applied in constructing confidence intervals (CIs) for A_{12}.

Since both *FP* and *FPC* have a test statistic of the form d/S_d, where d is an estimate of the stochastic difference and s_d is an estimator of its standard error, an $1 - \alpha -$ CI for δ looks like the following:

$$C_{1-\alpha} = (d \pm x_{crit} s_d), \tag{24}$$

where x_{crit} is the two-sided critical value corresponding to the α significance level. In the small samples case (if n and m are not greater than 12), x_{crit} can be determined by Table 1 of Fligner and Policello (1981) or Table 14 of Appendix B of Wilcox (1996). In other cases x_{crit} is the $1 - \alpha/2$ quantile of the standard normal distribution.

If one has a (c_L, c_U) CI for δ, one can very easily get one for A_{12} as well, applying the $A_{12} = (\delta + 1)/2$ linear transformation on the c_L, c_U limits (see (9)). We note that the same CI formula has been reported by Wilcox (1996, p. 371-372) for p_* in the continuous case, where $p_* = A_{12}$.

A CI based on *rW* can be constructed as follows. First remember that *rW* is the Welch test performed on the rank scores and hence

$$rW = (R_1/m - R_2/n)/D_{rW}, \tag{25}$$

where R_1 and R_2 are the rank sums of samples 1 and 2 computed as in the *MWW* test and

$$D_{rW} = [(SD_{R1})^2/m + (SD_{R2})^2/n]^{1/2}, \tag{26}$$

where SD_{R1} and SD_{R2} are the sample *SD*s of the rank scores of samples 1 and 2 respectively. Now from (14) and the analogous expression for A_{21} we have

$$R_1/m - R_2/n = n\hat{A}_{12} - m\hat{A}_{21} - (m - n)/2. \tag{27}$$

From identity (14) it is easily seen that analogously to the $A_{21} = 1 - A_{12}$ equality (see (3))

$$\hat{A}_{12} = 1 - \hat{A}_{21} \tag{28}$$

holds. Based on this we have for the difference of the rank means from (27)

$$R_1/m - R_2/n = (n + m)\hat{A}_{12} - (m + n)/2 = N(\hat{A}_{12} - .5). \tag{29}$$

Then we can write *rW* in the following form:

$$rW = (\hat{A}_{12} - .5)/(D_{rW}/\mathrm{N}). \tag{30}$$

Since *rW* can be approximated with a *t* distribution with an appropriate adjusted degrees of freedom, df_W (see, e.g., Zimmerman & Zumbo, 1993b, p. 507 or Welch, 1996, p. 133), which can be calculated from the rank sample variances and sample sizes, a $1 - \alpha$ confidence interval for A_{12} is

$$(c_L, c_U) = (\hat{A}_{12} \pm t(df_W, \alpha) D_{rW}/N). \tag{31}$$

Here $t(df_W, \alpha)$ is the two-sided critical value of Student's *t* distribution with df_W degrees of freedom at the α significance level, and $\hat{A}_{12}$ can be computed by formula (14).

Example 2

In an examination of admittance to the Psychology major at Eötvös Loránd University, Budapest, 1981, the number of males and females were $m = 16$ and $n = 78$ respectively. The means (and *SD*s) of the applicants' ages were 26.8 ($SD = 6.5$) and 21.7 ($SD = 5.3$) years for the male and female applicants respectively. In order to test the null hypothesis that male and female applicants are stochastically equal with respect to their age, a Mann-Whitney-Wilcoxon test was performed ($R_1 = 1087$, $R_2 = 3378$, $U = 3.34$, $p < .01$). Since the potentially different population variances may invalidate the *MWW* test, the three robust rank tests described in section 3 were also performed, yielding the following results:

rW: $t'_R = 3.995$ ($df_W = 25$, $p < .01$);

FP: $Z_{FP} = .524/.1428 = 3.669$ ($p < .01$);

FPC: $Z_{FPC} = .524/.1251 = 4.189$ ($p < .01$).

Having these highly significant results, the claim that male and female applicants are stochastically unequal with respect to their age, seems to be highly convincing.

In order to assess the extent to which these groups differ, first the point estimate for A_{12} is being computed by means of formula (14):

$$\hat{A}_{12} = (1087/16 - 17/2)/78 = .762.$$

This value reflects a substantial dominance of males' age values over that of females (see Table 1).

Using the observed *SD*s of the rank scores in the two samples (they happened to be 21.679 and 25.956, respectively), from formula (26) we get

$$D_{rW} = (21.679^2/16 + 25.956^2/78)^{1/2} = 6.1653.$$

And given the two-sided critical value in the table of the *t*-distribution for $df_W = 25$ at the .05 level is 2.060 formula (31) yields the following .95 CI for A_{12}:

$$(c_L, c_U) = (.762 \pm (2.06)(6.1653)/94) = (.627, .897).$$

Next we note that one can compute a .95 CI for δ (instead of for A_{12}) based on *FP* as follows:

$$C_{.95} = (.524 \pm (1.96)(.1428)) = (.244, .804),$$

or based on *FPC* as follows:

$$C_{.95} = (.524 \pm (1.96)(.1251)) = (.279, .769)$$

(see formula (24)). From these CIs, one can use formula (9) to compute the corresponding .95 CIs for A_{12} to compare with the CI for A_{12} derived above based on *rW*:

$$\text{via } FP\text{: } (c_L, c_U) = ((.244 + 1)/2, (.804 + 1)/2) = (.622, .902),$$
$$\text{via } FPC\text{: } (c_L, c_U) = ((.279 + 1)/2, (.769 + 1)/2) = (.640, .885).$$

Based on the guidelines of Table 1, these estimations of A_{12} and δ show a medium to large dominance of male applicants over female ones with respect to their age. The small inconsistencies between the obtained three CIs for A_{12} can be explained by the simulation results summarized in Table 4. In our example we have samples with slightly different *SD*s but with very different sample sizes. In this case *rW*'s Type I error rate is very close to its nominal level, whereas those for *FP*, and especially for *FPC* are duly inflated (see the second column in Table 4). For this reason the CI based on *rW* should be regarded as the most reliable. We note also that the slightly tighter CI for A_{12} based on *FPC* compared to that based on *FP* reflects the slightly more extreme inflation of *FPC* compared to that of *FP* under the given conditions (see Table 4).

5. The I-Group Case

CL and *A*, as effect size indicators, were introduced for the purpose of measuring the extent to which two populations differ from each other with respect to a variable *X*. It is therefore desirable to have a unique baseline value (such as zero) in the case when the two populations have identical distributions. This baseline is zero if one uses Cohen's effect size coefficients, and .5 with *CL* and *A*, but these latter measures can be simply transformed (by a subtraction of .5) to yield the same zero value effect size in the case of identical distributions.

It is highly desirable to maintain this feature when one generalizes these effect size measures to the multi-group case. However, the solution proposed by McGraw and Wong (1992) fails to fulfill this expectation. To measure the extent to which distribution 1 differs from the others together in the *I*-group case, McGraw and Wong proposed as a generalization of *CL* "the probability that a score from group A will be simultaneously larger than scores sampled from the other groups (B, C, etc.)" (1992, p. 362-363); that is:

$$CLg = P(X_1 > X_2, X_1 > X_3, \ldots, X_1 > X_I). \tag{32}$$

Since according to McGraw and Wong, "there is no exact method" of determining the above probability (1992, p. 363), they proposed the following method to estimate *CLg*. Let us compute first the product of the component probabilities:

$$p = P(X_1 > X_2)P(X_1 > X_3) \ldots P(X_1 > X_I).$$

Then estimate *CLg* via p by empirically derived linear regression equations which utilize the large positive correlation between *CLg* and p. These empirically derived linear regression equations have the form of

$$p^* = Cp + D, \tag{33}$$

and are $p^* = .88p + .11$, $p^* = .90p + .15$, and $p^* = .93p + .17$ in the three-, four-, and five-group cases respectively (see Table 3 of McGraw & Wong, 1992). In these equations p^* denotes the estimate of *CLg*. In order to show what strange *CLg* values can be obtained by using this method, suppose that the distributions to be compared are all identical, i.e., that there is no difference between the populations to be compared. What will McGraw and Wong's proposed *CLg* measure yield in this case? If variable X is continuous, and we have independent samples, $P(X_1 > X_i)$ will equal .5 for any random score X_1 from sample 1 and random score X_i from sample i ($i = 1, \ldots, I$). For this reason in the three-group case the p coefficient in equation (33) will be

$$p = P(X_1 > X_2)P(X_1 > X_3) = (.5)(.5) = .25,$$

which yields the following estimated *CLg* value:

$$p^* = (.88)(.25) + .11 = .33.$$

Similarly in the four-group case $p = (.5)(.5)(.5) = .125$ and $p^* = .90p + .15 = .2625$, and in the five-group case $p = (.5)(.5)(.5)(.5) = .0625$ and $p^* = .93p + .17 = .2281$.

The conclusion is that McGraw and Wong's proposed generalization of their effect size measure does not have a unique constant value that would indicate the "no difference between the distributions" state, independently of the number of groups to be compared.

Another deficiency of their proposed generalization is that *CLg* can only measure a deviation of each population from the others together, but cannot provide a single value indicating the overall extent of difference among the I populations. For this reason, in the I-group case one has to compute I different *CLg* measures, not just one.

Our main idea for generalizing A to the I-group case consists first of generalizing the concept of stochastic equality from the two-group case to the multi-

group case, and then measuring the extent to which the populations to be compared are stochastically unequal with respect to variable X.

For the definition of a generalized stochastic equality we have two candidates:

D1. *Stochastic homogeneity*, when any one of the I populations is stochastically equal to the union of the remaining $I - 1$ populations:

$$A_{iu} = .5 \text{ for all } i \ (i = 1, \ldots, I).$$

If there is a set of r_i proportions ($i = 1, \ldots, I$, $\Sigma r_i = 1$) assigned to the I populations, and the unification of the remaining $I - 1$ populations is always processed with these relative proportions as weights, then it can be derived algebraically that

$$A_{iu} = \Sigma_{k \neq i} w_k A_{ik}, \tag{34}$$

where

$$w_k = r_k/(1 - r_i). \tag{35}$$

With a short derivation it can also be seen that any one of the I populations will stochastically equal the union of the remaining $I - 1$ populations if and only if this population is stochastically equal to the union of all I populations (in the same way that any number out of a set of numbers will equal the average of the whole set if and only if it equals the average of the remaining ones).

The choice of the r_i proportions may depend on the research question. If the populations to be compared can be regarded as subpopulations of a larger common population with known proportions, these proportions can serve as the r_i weights. For example, such differential weighting might be appropriate in a study of racial differences when a simple random sample is drawn from a population composed of racial groups of different sizes. However, in certain cases identical weights ($r_1 = r_2 = \ldots = r_I = 1/I$) may be preferable. As an example suppose that we want to compare different treatments or experimental conditions with independent samples. In this case an unequal weighting of the different populations (distributions) representing the different treatments or conditions can hardly be justified.

D2. *Pairwise stochastic equality*, when any two of the I populations are stochastically equal, i.e., when

$$A_{ik} = .5 \text{ for all } (i, k) \text{ pairs.}$$

Both for D1 and D2 the extent to which the I populations are heterogeneous will depend on how much the A_{iu} or A_{ik} values differ from .5. Accordingly, a measure of a global stochastic heterogeneity should accumulate the individual deviations from the .5 value. One such measure can be for D1 the average absolute deviation from .5:

$$AAD = \Sigma_i |A_{iu} - .5| / I, \tag{36}$$

and for D2 the average absolute pairwise deviation from .5:

$$AAPD = (\Sigma_{i<k} |A_{ik} - .5|)/(I(I-1)/2). \tag{37}$$

It is readily seen that in the $I = 2$ case both *AAD* and *AAPD* reduce to $|A_{12} - .5|$, which also equals $|A_{21} - .5|$. For this reason the values of *AAD* and *AAPD* can be evaluated according to Table 1 if we add .5 to them. Note that in contrast to A_{12}, which only reflects the extent of superiority of population 1 over population 2, these measures reflect a summary of the extent of differences among the individual populations. The direct generalization of the directional A_{12} measure is the set of A_{iu} components in (36), which obviously reduces to $A_{1u} = A_{12}$ and $A_{2u} = A_{21}$ in the $I = 2$ case.

The *AAD* effect measure equals 0 if and only if all A_{iu} ($i = 1, \ldots, I$) measures equal .5, i.e., if and only if each population is stochastically equal to the weighted union of the rest. For the designation of this zero effect case we used the term "stochastic homogeneity" (Vargha & Delaney, 1998). Stochastic homogeneity always holds when the A_{ik} pairwise stochastic superiority values are all equal (in this case the common value is necessarily .5 and $AAPD = 0$), but the reverse implication is not true, that is $AAD = 0$ does not imply $AAPD = 0$ (for more details see Vargha & Delaney, 1998).

An alternative approach may be the use of the squared deviations instead of the absolute deviations, like in the formula of the standard deviation and variance. The basic difference between the two types of approach is that the absolute deviation approach assigns equal weights to the different deviations, whereas the squared deviation approach assigns to a larger deviation larger weight than to a smaller one. Due to this the square-root of the mean squared deviation, a kind of *SD*, is generally slightly greater than the corresponding *AAD* measure.

Estimation of AAD

First of all we have to decide what kind of weighting we want to use in the formulas for the A_{iu} components (see (34) and (35)). If the populations to be compared can be regarded as the subpopulations of a common larger population, then the $(r_1, r_2, \ldots, r_I)$ set of relative sizes of the subpopulations can be a reasonable choice. But an important aspect during the estimation of *AAD* is that we must insure the applied sample sizes be proportional to the corresponding population sizes. If the r_i proportions are known constants, then the sample sizes can be determined so they fulfill the

$$n_i/N = r_i \ (i = 1, \ldots, I) \tag{38}$$

condition. If the subpopulation sizes (or proportions) are unknown, a large random selection from the common united population may ensure at least approximately that (38) be fulfilled. If (38) holds then for (35) one gets

$$w_k = (n_k/N)/(1 - n_i/N) = n_k/(N - n_i), \tag{39}$$

and hence in this case (34) can be expressed as follows:

$$A_{iu} = \Sigma_{k \neq i} n_k A_{ik}/(N - n_i). \tag{40}$$

Now a point estimation for *AAD* can be obtained as follows. Draw a random sample of X values from each population i of size n_i, independently from each other. Then rank the whole sample of size N together and denote the sum of the ranks in group i by R_i. It can be proven that

$$E(R_i/n_i) = (n_i + 1)/2 + \Sigma_{k \neq i} n_k A_{ik} \tag{41}$$

(see identity (18) in Vargha & Delaney, 1998). From (40) and (41) it follows directly that

$$E(R_i/n_i) = (n_i + 1)/2 + (N - n_i)A_{iu}, \tag{42}$$

and hence

$$A_{iu} = [E(R_i/n_i) - (n_i + 1)/2]/(N - n_i). \tag{43}$$

From (43) we can now conclude that A_{iu} can be estimated via the R_i rank sums as follows:

$$\hat{A}_{iu} = [R_i/n_i - (n_i + 1)/2]/(N - n_i). \tag{44}$$

Therefore inserting these $\hat{A}_{iu}$ estimates into the right side of (36) we will obtain an unbiased point estimate of *AAD*. We note that the R_i rank sums are readily available in statistical program packages that contain nonparametric group comparison procedures (such as the Kruskal-Wallis test).

Estimation of AAPD

Again draw a random sample of X values from each population i of size n_i, independently from each other. Then estimate all pairwise A_{ik} stochastic superiority values ($i = 1, \ldots, I - 1$; $k = i + 1, \ldots, I$) by a joint ranking of the $n_i + n_k$ observations from samples i and k alone, applying expression (14) for computing $\hat{A}_{ik}$. Inserting these $\hat{A}_{ik}$ estimates into (37) we get a point estimate for *AAPD*.

Significance of AAD

It can be proven by a mathematical derivation that for any variable X of at least ordinal type the identity

$$A_{iu} = .5 \text{ for all i } (i = 1, \ldots, I) \tag{45}$$

is equivalent to the following one:

$$E(R_1/n_1) = E(R_2/n_2) = \ldots = E(R_I/n_I) \tag{46}$$

(for the proof see Vargha & Delaney, 1998). For this reason the $AAD = 0$ hypothesis, which states that the populations to be compared are stochastically homogeneous with respect to X, can be tested by a one-way ANOVA performed on the rank scores, which we call rank ANOVA. It can easily be verified that this method is equivalent to the large sample procedure of the Kruskal-Wallis test. Furthermore, if the variance homogeneity assumption of the rank ANOVA is not tenable, a robust alternative to the ANOVA (such as the James test or the Welch test; see Wilcox, 1996, p. 182 and p.183) on the rank scores can be applied.

Remember that the proportion of the sample sizes must equal at least approximately the proportion of the population sizes. This condition ensures that the Kruskal-Wallis test will be valid in testing the null hypothesis of stochastic homogeneity with the population sizes serving as the population weights in (35) (see also (34) and (38)). This restriction is due to the consistency condition of the Kruskal-Wallis test, and therefore it cannot be overlooked (see, Kruskal, 1952, p. 533 for continuous and p. 537 for discrete distributions; Noether, 1967, p. 51-52). Accordingly, if one does not want the sample sizes to enter the formulation of null hypothesis, one must use equal samples, and identical r_k and w_k weights in (34) and (35).

Significance of AAPD

Fligner (1985) introduced a new test of the null hypothesis of identical continuous distributions versus the alternatives of the form

$$H_1: P(X_i < X_k) - .5 \neq 0,$$

i.e., versus pairwise stochastic inequalities. Fligner's test statistic, W_N, is defined as

$$W_N = \frac{12}{N+1} \Sigma_{i<k} n_i n_k (W_{ik})^2, \tag{47}$$

where

$$W_{ik} = [R_{ik} - .5n_i(n_i + n_k + 1)]/(n_i n_k). \tag{48}$$

Here R_{ik} denotes the sum of the ranks associated with the i-th sample in the joint ranking of the n_i+n_k observations from samples i and k alone (the .5 multiplier in (48) is incorrectly missing from the formula given in Fligner's paper). If H_0 is true and the $n_1, n_2, \ldots, n_I$ sample sizes are not very small, then W_N follows approximately a chi-square distribution with $df = I - 1$ degrees of freedom, just

as the H test statistic of the Kruskal-Wallis test. Fligner (1985) provided some theoretical evidence that his test performs at least as well as the Kruskal-Wallis test. But this latter has special tables with exact tail probabilities only with sample sizes not exceeding five (see Hollander & Wolfe, 1973). Accordingly the Fligner test with the chi-square approximation may also perform well with $n > 5$ for each sample. Certainly this claim should be confirmed empirically by appropriate simulation studies.

It can be proven by a short algebraic derivation that $W_{ik} = \hat{A}_{ik} - .5$, where $\hat{A}_{ik}$, the estimated pairwise stochastic superiority measure for populations i and k, is defined by formula (14). Accordingly, we have for W_N:

$$W_N = \frac{12}{N+1} \Sigma_{i<k} n_i n_k (\hat{A}_{ik} - .5)^2. \tag{49}$$

From (49) it follows that W_N is sensitive to any pairwise stochastic inequality of the populations to be compared, and thus it can be used for testing the $AAPD = 0$ hypothesis. Because the W_N test statistic given by (47) is very similar to H (the test statistic of the Kruskal-Wallis test) if Fligner's test is used to test the $AAPD = 0$ hypothesis, the assumption of variance homogeneity may well be required (see Vargha & Delaney, 1998).

Example 3

From an archival data set including 807 Rorschach protocols that served as the basis for the construction of the Hungarian Rorschach Standard (Vargha, 1989), three random samples of size 79 each were selected, representing three different levels of education (low, medium, and high). To compare the level of the Positive Human Movement percentage index (H+%) across the three education levels, a Kruskal-Wallis test was performed ($H = 6.90$, $df = 2$, $p < .05$). The rank means for the three samples happened to be 119.80, 104.34, and 132.85, respectively. Thus the estimated A_{iu} values based on formula (44) are as follows:

$$\hat{A}_{1u} = (119.80 - 80/2)/158 = .5051,$$
$$\hat{A}_{2u} = (104.34 - 80/2)/158 = .4072,$$
$$\hat{A}_{3u} = (132.85 - 80/2)/158 = .5877.$$

Finally, inserting these values into formula (36) we get a value of .062 for an unbiased point estimate of AAD. Since $AAD + .5 = .562$, AAD reflects here a small amount of stochastic heterogeneity (see Table 1).

With the same data we calculated the estimates of the pairwise stochastic superiority measures and obtained the following results:

$$\hat{A}_{12} = .5708, \hat{A}_{13} = .4398, \text{ and } \hat{A}_{23} = .3853.$$

Inserting these values into formula (37) we get a value of .082 for an estimate of *AAPD*. Since $AAPD + .5 = .582$, *AAPD* reflects a small extent of pairwise stochastic heterogeneity (see Table 1).

With these $\hat{A}_{ik}$ values we can also compute the W_N test statistic of Fligner's test as follows:

$$W_N = (12/237)79^2(.0708^2 + .0602^2 + .1147^2) = 6.89,$$

which is significant at the same level ($p < .05$), as the Kruskal-Wallis test above.

6. The Correlated Samples Case

The approach proposed by McGraw and Wong (1992) for assessing the difference between two correlated samples suffers from the same deficiency as the one proposed to compare independent samples. It requires not only the continuity of the distribution of the dependent variable but also the normality. We are now going to introduce a generalization of *CL* for this situation and a method of estimation, which does not need such a restrictive assumption. Let us first define the measure of stochastic superiority of a variable X over a variable Y by formula

$$A_{XY} = P(X > Y) + .5P(X = Y). \tag{50}$$

With any random sample of size n of the (X,Y)-values, A_{XY} can be estimated very easily as follows:

$$\hat{A}_{XY} = [\#(X > Y) + .5\#(X = Y)]/n. \tag{51}$$

It is readily seen that A_{XY} can be defined for any X and Y variables of at least ordinal type, and that it equals *CL* when both X and Y are continuous. If we again define stochastic equality by the $A_{XY} = .5$ equality, it is easy to see that this stochastic equality is equivalent to the identity

$$P(X > Y) = P(X < Y). \tag{52}$$

For this reason the stochastic equality of variables X and Y can be tested by means of the well known sign test (see, e.g., Maritz, 1995, p. 24). We note that in the continuous case the stochastic equality of X and Y is equivalent as well to saying the median of $X - Y$ is 0.

For a simple way of finding confidence limits for A_{XY} in the continuous case see Maritz (1995, p. 25).

Generalization to the I *Matched Samples Case*

Our suggestion to generalize A_{XY} to the I matched samples case is analogous to the presented I-group generalization of *CL*. The definition and point estimation of *AAD* (the average absolute deviation from .5), and *AAPD* (the average

absolute pairwise deviation from .5) is very simple and straightforward. However, as we did not find a method for testing the $AAPD = 0$ hypothesis, only details with respect to *AAD* will be provided, which can be summarized as follows:

(i) Determine the pairwise $A_{ik} = A(X_i, X_k)$ stochastic superiority values for all (i,k) pairs of indexes $(i = 1, \ldots, I; k = 1, \ldots, I)$;

(ii) Calculate for each variable X_i $(i = 1, \ldots, I)$ the

$$\mathrm{A}_{iu} = \Sigma_{k \neq i} \mathrm{A}_{ik}/(I - 1) \tag{53}$$

measure. A_{iu} is the average of the stochastic superiority measures when variable i is compared in turn to each of the others;

(iii) Compute *AAD* for these A_{iu} measures in the same way as was indicated in (36).This *AAD* effect measure equals 0 if and only if all A_{iu} $(i = 1, \ldots, I)$ values equal .5. To designate this zero effect case we can say that variables X_i $(i = 1, \ldots, I)$ are "stochastically homogeneous."

Estimation of AAD

Similar to the two matched samples case, draw a random $(X_1, X_2, \ldots, X_I)$ sample of size n. Then rank the whole sample case-wise, as is done in the Friedman test (see, e.g, Maritz, 1995, p. 191). It can be proven fairly simply that

$$E(R_i/n_i) = 1 + \Sigma_{k \neq i} \mathrm{A}_{ik} = 1 + (I - 1)\mathrm{A}_{iu}, \tag{54}$$

from where

$$\mathrm{A}_{iu} = [E(R_i/n) - 1]/(I - 1). \tag{55}$$

From (55) we can now conclude that A_{iu} can be estimated via the R_i rank sums as follows:

$$\hat{A}_{iu} = [(R_i/n) - 1]/(I - 1). \tag{56}$$

Therefore, by inserting these $\hat{A}_{iu}$ estimates into the right side of (36) we will obtain a point estimate of *AAD*. We note that the R_i rank sums are available in statistical program packages that contain the Friedman test.

Significance of AAD

It is easily seen from (55) that in the I matched samples case the identity

$$\mathrm{A}_{iu} = .5 \text{ for all } i \ (i = 1, \ldots, I) \tag{57}$$

is equivalent to the following one:

$$E(R_1/n) = E(R_2/n) = \ldots = E(R_I/n). \tag{58}$$

For this reason the $AAD = 0$ hypothesis, which states that the variables to be compared are stochastically homogeneous, can be tested by a one-way repeated sample ANOVA performed on the rank scores. It can easily be verified that this method is equivalent to the large sample procedure of the Friedman test (see Winer, 1971, p. 301–302).

Example 4

In a psychological experiment the Ss were 91 male athletes whose pulse rates were measured after a two-hour-long practice. In this experiment the pulse rate was measured three times: first at the beginning of the experiment (PR1), second after a frustrating psychological intervention (PR2), and third at the end of the experiment (PR3). To test the null hypothesis of stochastic homogeneity of variables PR1, PR2, and PR3, the Friedman-test was performed ($G = 25.04$, $df = 2$, $p < .01$). The rank sums for the three variables happened to be 170, 220, and 156 respectively. From formula (56) we now get for the three estimated A_{iu} values as follows:

$$\hat{A}_{1u} = (170/91 - 1)/2 = .4341,$$
$$\hat{A}_{2u} = (220/91 - 1)/2 = .7088,$$
$$\hat{A}_{3u} = (156/91 - 1)/2 = .3571.$$

Finally, inserting these values into formula (36) we get a value of .139 for the point estimation of *AAD*. Since $AAD + .5 = .639$, *AAD* reflects now a medium level of stochastic heterogeneity (see Table 1).

Discussion

The *CL* effect size indicator proposed by McGraw and Wong (1992) expresses the difference between two populations with regard to a variable in terms of a probability. The only requirement imposed upon this variable is that the values from the two populations should be able to be compared with each other to determine which of them is greater, which of them is smaller. This requirement is fulfilled if the variable can be measured on an ordinal scale, but apparently McGraw and Wong did not realize this. For this reason they presumed it was necessary to have the variable satisfy some of the basic requirements of the normal distribution, namely continuity, and the computability of the mean and the standard deviation.

Relying only on ordinal relationships leads to the notions of measures of stochastic superiority, stochastic equality, and stochastic homogeneity as well as the well-known nonparametric comparison tests. Major contributions to this topic have been recently published by Vargha and Delaney (1998). These new contributions made it possible to introduce in this paper a mathematically correct generalization of *CL*, which we designate the measure of stochastic superiority, *A*. However, *A* is not without antecedents in the statistical literature. We showed that *A* is a simple linear transformation of the δ measure of

stochastic difference, detailed in Cliff (1993). The only difference between A and δ is that whereas A measures the stochastic difference of two distributions in the [0, 1] interval, δ measures it in the $[-1, 1]$ interval. It must be emphasized that A and δ apply equally well to both continuous and discrete parent distributions.

In our paper we have devoted much space to the notion of stochastic equality, which can be defined by either of the equivalent equalities, $A = .5$ or $\delta = 0$. We described three robust tests of stochastic equality: the rank Welch test (see Zimmerman & Zumbo, 1992, 1993a, 1993b); the Fligner-Policello test (Fligner & Policello, 1981); and Cliff's modification of the Fligner-Policello test (Cliff, 1993). With a small Monte Carlo analysis we provided some evidence that these methods are substantially more robust to variance heterogeneity than the rank t test, which is essentially the same as the Mann-Whitney-Wilcoxon test.

In the multi-group case we introduced the *AAD* and *AAPD* measures for assessing the extent of stochastic heterogeneity of more than two distributions and provided details concerning their point estimations and significance tests for both the independent and correlated samples cases.

Our important contribution related to *CL* and A is the concept of stochastic equality with which we designate the zero effect case of this stochastic comparison (i.e., when $A = .5$). The relevance of this concept is well reflected in the numerous recently published papers dealing with two-group comparison tests that are appropriate for testing stochastic equality (Fligner & Policello, 1981; Zimmerman & Zumbo, 1992, 1993, 1993b; Cliff, 1993; Zumbo & Coulombe, 1997).

We generalized A as well as the concept of stochastic equality for the multi-group case too. However, the presented methods are sometimes tentative. More empirical evidence is needed to identify the best of the possible competitive measures and empirical studies concerning the properties of the proposed significance tests would be informative as well.

References

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494–509.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Fligner, M. A. (1985). Pairwise versus joint ranking: Another look at the Kruskal-Wallis statistic. *Biometrika, 72*, 705–709.

Fligner, M. A., & Policello, II, G. E. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association, 76*, 323–327.

Friedman, H. (1968). Magnitude of experimental effect and a table of its rapid estimation. *Psychological Bulletin, 70*, 245–251.

Gibbons, J. D. & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed.). New York: Marcel Dekker.

Hájek, J. (1969). *A course in nonparametric statistics*. San Francisco: Holden-Day.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. New York: Wiley.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: Wiley.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics, Vol. 2: Inference and relationship* (3rd ed.). London: Griffin.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics, 23*, 525–540.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. San Franciso: Holden-Day.

Levy, P. (1967). Substantive significance of significant differences between groups. *Psychological Bulletin, 67*, 37–40.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*, 50–60.

Maritz, J. S. (1995). *Distribution-free statistical methods* (2nd ed.). London, New York, Tokyo: Chapman & Hall.

Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data. A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361–365.

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

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 539–543.

Miller, R. G. (1986). *Beyond ANOVA. Basics of applied statistics*. New York: Wiley.

Noether, G. E. (1967). *Elements of nonparametric statistics*. New York: Wiley.

Onghena, P. (1993). A theoretical and empirical comparison of mainframe, microcomputer, and pocket calculator pseudorandom number generators. *Behavior Research Methods, Instruments, & Computers, 25*, 384–395.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., & Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics, 21*, 201–209.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. New York: Wiley.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166–169.

Vargha, A. (1989). *A Magyar Rorschach Standard táblázatai*. [The tables of the Hungarian Rorschach Standard.] Budapest: Schoolbook Publisher.

Vargha, A. (1999). *MiniStat 3.1 verzió. Felhasználói kézikönyv*. [MiniStat, Version 3.1. Manual] Budapest: Pólya Kiadó.

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics, 23*, 170–192.

Wilcox, R. R. (1990). Determining whether an experimental group is stochastically larger than a control. *British Journal of Mathematical and Statistical Psychology, 43*, 327–333.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, New York: Academic Press.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). Tokyo, London: McGraw-Hill Kogakusha.

Zimmerman, D. W., & Zumbo, B. D. (1992). Parametric alternatives to the Student t test under violation of normality and homogeneity of variance. *Perceptual and Motor Skills, 74*, 835–844.

Zimmerman, D. W., & Zumbo, B. D. (1993a). Rank transformations and 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.

Zimmerman, D. W., & Zumbo, B. D. (1993b). The relative power of parametric and nonparametric statistical methods. In: G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 481–517). Hillsdale, NJ: Erlbaum.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology, 51*, 139–149.

Authors

ANDRÁS VARGHA is Associate Professor, Department of Experimental Psychology, Eötvös Loránd University, Izabella utca 46, H–1064 Budapest, Hungary; vargha@izabell.elte.hu. He specializes in nonparametric comparison methods and investigations of the robustness of classical statistical methods to violations of their assumptions.

HAROLD D. DELANEY is Professor and Associate Chair, Department of Psychology, University of New Mexico, Albuquerque, NM 87131; hdelaney@unm.edu. He specializes in applied statistics and individual differences.