Multivariate Behavioral Research, 47:201–223, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2012.658329

Psychology Press
Taylor & Francis Group

Confidence Intervals for the Probability of Superiority Effect Size Measure and the Area Under a Receiver Operating Characteristic Curve

John Ruscio and Tara Mullen
The College of New Jersey

It is good scientific practice to the report an appropriate estimate of effect size and a confidence interval (CI) to indicate the precision with which a population effect was estimated. For comparisons of 2 independent groups, a probability-based effect size estimator (A) that is equal to the area under a receiver operating characteristic curve and closely related to the popular Wilcoxon-Mann-Whitney nonparametric statistical tests has many appealing properties (e.g., easy to understand, robust to violations of parametric assumptions, insensitive to outliers). We performed a simulation study to compare 9 analytic and 3 empirical (bootstrap) methods for constructing a CI for A that can yield very different CIs for the same data. The experimental design crossed 6 factors to yield a total of 324 cells representing challenging but realistic data conditions. Results were examined using several criteria, with emphasis placed on the extent to which observed CI coverage probabilities approximated nominal levels. Based on the simulation study results, the bias-corrected and accelerated bootstrap method is recommended for constructing a CI for the A statistic; bootstrap methods also provided the least biased and most accurate standard error of A. An empirical illustration examining score differences on a citation-based index of scholarly impact across faculty at low-ranked versus high-ranked research universities underscores the importance of choosing an appropriate CI method.

Correspondence concerning this article should be addressed to John Ruscio, The College of New Jersey, P.O. Box 7718, Ewing, NJ 08628. E-mail: ruscio@tcnj.edu

CONFIDENCE INTERVALS FOR THE PROBABILITY OF SUPERIORITY WHEN COMPARING TWO INDEPENDENT GROUPS

According to the American Psychological Association's (2009) *Publication Manual*, it is almost always necessary to report an appropriate estimate of effect size in the Results section of a publication in order for the reader to fully understand the magnitude of the effect and properly contextualize the importance of the findings. The use of effect size estimates in published studies is not required by all journals, but it does serve as a useful adjunct to the standard reporting of statistical significance. Stating that there is a statistically significant effect implies that there is an effect beyond what would be expected due to sampling error. A well-chosen effect size estimator can be used to characterize the magnitude of the effect and help readers understand the practical significance of results (Wilkinson & the APA Task Force on Statistical Inference, 1999). In addition, the *Publication Manual* recommends reporting a confidence interval (CI) to provide information on the precision with which the population effect size has been estimated. This can be useful for interpreting results within a study and for comparing results across studies (Thompson, 2002; Wilkinson et al., 1999).

When comparing two independent groups, an effect size index A that estimates what Grissom and Kim (2005) call the "probability of superiority" (p. 98) has a number of desirable characteristics (Ruscio, 2008). Unlike conventional statistics that are not robust to violations of their parametric assumptions and can be difficult to interpret without statistical expertise, the nonparametric effect size index A simply estimates the probability that a member of one population scores higher than a member of another population. After reviewing important differences between A and conventional effect size indices—plus connections between A, nonparametric test statistics, and the area under a receiver operating characteristic (ROC) curve—we expand upon and evaluate methods for constructing CIs for the versatile A statistic.

CONVENTIONAL EFFECT SIZE ESTIMATORS FOR TWO INDEPENDENT GROUPS

When comparing two independent groups, researchers can describe the magnitude of an effect in several ways. To help illustrate each of these, we introduce a small sample of hypothetical data ($n_x = n_y = 15$) representing health ratings for members of randomly assigned treatment (x) and control (y) groups; higher ratings indicate better health:

$$x = \{6, 7, 8, 7, 9, 6, 5, 4, 7, 8, 7, 6, 9, 5, 4\}$$

$$y = \{4, 3, 5, 3, 6, 2, 2, 1, 6, 7, 4, 3, 2, 4, 3\}$$

The most commonly used effect size estimator for data like these is Cohen's *d*, the standardized mean difference (Cohen, 1988). This is calculated as the difference between the group *M*s divided by the within-group *SD*, and it estimates the standardized mean difference between the populations from which the two samples were drawn.[1] Because this is standardized, results for different variables within a study or the same variable measured in different studies can be compared even when their scales differ. Weaknesses include the fact that *d* is sensitive to unequal group sizes when unequal *SD*s are pooled across groups, nonrobust to outliers, and can be difficult to understand for those untrained in statistics (McGrath & Meyer, 2006; Ruscio, 2008). For our illustrative data, $d = 1.728$. This is clearly a very large effect, but it may not be clear to laypersons to say that members of the treatment group members scored 1.728 *SD* units higher than members of the control group.

Another way to estimate the size of an effect is the point-biserial correlation (r_{pb}). This is calculated as the correlation between group membership (coded using any two unique values) and the dependent variable, and it estimates the corresponding population correlation. Like Cohen's d, r_{pb} is also standardized. When squared, it represents the proportion of variance in the dependent variable that can be explained by group membership. In addition to sharing many of the limitations of *d*, such as being nonrobust to outliers and nonintuitive for laypersons, r_{pb} is especially sensitive to the relative sizes of the two groups. The extent and implications of this heightened sensitivity to group sizes are discussed and illustrated by McGrath and Meyer (2006) and Ruscio (2008). For our illustrative data, $r_{pb} = .667$. Many people would find it difficult to grasp what it means to say that group membership correlated $r_{pb} = .667$ with health ratings (or that it explains 44.4% of the variance in health ratings; $.667^2 = .444$).

ESTIMATING THE PROBABILITY OF SUPERIORITY

An alternative to indices such as *d* and r_{pb} is to express the size of an effect using a statistic that estimates the probability that a randomly selected member of population *X* scores higher than a randomly selected member of population Y: $\Delta = Pr(X > Y)$. In contrast to a comparison of means or another location measure, this is conventionally referred to as stochastic superiority (e.g., Vargha & Delaney, 2000) or the probability of superiority (e.g., Grissom & Kim, 2005). An estimate of Δ may be easier to understand than *d* or r_{pb}, especially for those with little or no statistical expertise (Hsu, 2004). For example, rather than estimating a health benefit in within-group *SD* units or as a correlation

[1] Closely related estimates are Glass's (1976) Δ and Hedges's (1981) g, which differ from Cohen's d in the denominator used to standardize the mean difference between groups. For an excellent overview of these and other effect size estimators, see Kirk (1996).

with group membership, one can estimate the probability of better health with treatment than without it. In addition to being a more intuitive statistic than d or r_{pb}, the nonparametric estimate of Δ described later is not sensitive to group sizes and is much more robust to unequal variances or outliers (Ruscio, 2008).

Two distinct ways of estimating Δ have been described. First, Wolfe and Hogg (1971) introduced a statistic that McGraw and Wong (1992) later popularized using the label of "common language effect size" (CL). This is calculated using standard parametric assumptions of population normality and equal variances, and in the case of equal-size groups ($n = n_x = n_y$), CL can be expressed in terms of the usual t value for a comparison of two independent groups as $CL = \Phi(\frac{t}{\sqrt{n}})$, where Φ is the normal cumulative distribution function (i.e., $\Phi[z_\alpha] = \alpha$). For our illustrative data, $M_x - M_y = 2.867$ and $s_{M_x - M_y} = 1.651$, so $t = 4.732$; with $n = 15$, $CL = \Phi(\frac{4.732}{\sqrt{15}}) = .889$. Even those untrained in statistics should have a fairly easy time understanding what it means to say that there is an 88.9% chance that the health rating would be higher for a randomly chosen member of the treatment group than for a randomly chosen member of the control group.

Whereas CL retains the parametric assumptions that render d and r_{pb} sensitive to outliers, subsequent investigators provided a nonparametric estimator of Δ that addresses these concerns. Delaney and Vargha (2002) expressed it in this form:

$$A = [\#(x > y) + .5\#(x = y)]/n_x n_y, \tag{1}$$

where # is the count function, x and y are vectors of scores for the two groups. Scores are compared across groups in all pairwise combinations, and ties are accommodated by assigning half credit. Here's how A would be calculated for our illustrative data. Beginning with $x_1 = 6$ and comparing this to each value of y yields 12 instances in which $x_1 > y_i$, 2 ties, and 1 instance in which $y_i > x_1$, so the numerator of A begins at $12 + .5(2) = 13$. For $x_2 = 7$, this adds 14.5 to the numerator of A (14 instances in which $x_2 > y_i$ plus .5 for 1 tie). Continuing through $x_{15} = 4$, the numerator of A sums to 199. The denominator of A is $n_x n_y = 15 \times 15 = 225$, so $A = 199/225 = .884$. The slight difference between the estimates $CL = .889$ and $A = .884$ suggests that the parametric assumptions imposed by CL are not entirely satisfied. When these assumptions are satisfied, CL will equal A. When they are not, we prefer using A, which does not require these assumptions and is therefore a more robust estimator of Δ. From this point forward, we focus exclusively on the use of A as an estimator of Δ.

The A statistic is closely related to several other statistics that require only ordinal data, such as the familiar Wilcoxon Rank Sum and Mann-Whitney U nonparametric test statistics (Delaney & Vargha, 2002; Fagerland & Sandvik, 2009; Zhou, 2008). Whereas these statistics are commonly used to test null hypotheses, neither they nor the A statistic are used frequently as an effect size estimator. Fortunately, readily available software can be used to obtain A with either of these procedures. SPSS calculates Mann-Whitney $U = \#(Y_1 < Y_2) + .5\#(Y_1 = Y_2)$, in which case A can be calculated as

$$A = \frac{n_x n_y - U}{n_x n_y}. \tag{2}$$

SPSS also reports the Wilcoxon test statistic as W_m, which can be converted to U (for use in Equation 2) as follows: $U = W_m - [n_s(n_s + 1)]/2$, where n_s is the smaller of the two sample sizes. Provided that one verifies how they are calculated, nonparametric test statistics from other software can be used. For example, the R function for the Wilcoxon test reports $W = n_x n_y - U$, which equals the numerator in Equation 2.

In addition to its relation to nonparametric test statistics, A is equal to a key statistic in signal detection theory (Swets, 1988; Swets, Dawes, & Monahan, 2000). Specifically, one can construct an ROC curve and obtain the area under the curve (*AUC*) as a measure of accuracy that is independent of the decision threshold (Fawcett, 2006). An ROC curve is plotted within a unit square as the relationship between the true positive rate (sensitivity) and false positive rate (1 − specificity) with which members of two groups are distinguished using one or more thresholds. For example, the graph on the left in Figure 1 shows the ROC curve for the illustrative data presented earlier. The graph on the right in Figure 1 shows that *AUC* can be calculated as the sum of the areas of the seven numbered trapezoids. When calculated using the trapezoidal method, $A = AUC$ (Hanley & McNeil, 1982); the SPSS module for ROC analysis provides *AUC* calculated in this way.

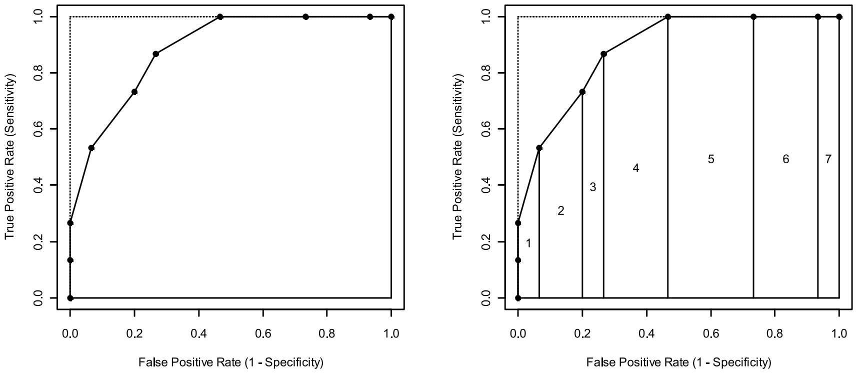

FIGURE 1 Receiver operating characteristic (ROC) curve for the illustrative data. The graph on the left shows the ROC curve, and the graph on the right shows that the area under the curve can be calculated as the sum of the areas of seven numbered trapezoids.

CONSTRUCTING CONFIDENCE INTERVALS FOR *A*

The attractive features of *A* have been described elsewhere (e.g., Ruscio, 2008), and the focus of the present research is on methods for constructing a CI for *A*. As noted earlier, the APA *Publication Manual* (2009) recommends providing CIs to indicate the precision with which parameters have been estimated. Prior research has introduced and examined many techniques for constructing a CI for *A*, but we believe there remains room for improvement because considerably greater emphasis has been placed on analytic CI methods rather than more computationally intensive, empirical CI methods. Some analytic approaches first require the calculation of an *SE* that is then used to construct a CI, whereas others use alternative methods that do not involve an *SE*. There are at least nine different analytic approaches to construct a CI for *A*,[2] all but two of which yield an interval symmetric about *A*. In other words, most analytic methods deal with sampling error by adding and subtracting the same amount to *A*, which yields a symmetric CI. Symmetric CIs presume symmetric sampling distributions, so one might expect the performance of these analytic methods to degrade as the sampling distribution of *A* deviates from symmetry. This sampling distribution will be symmetric only when $A = .50$, and it will become increasingly skewed as *A* departs from .50. Two analytic methods use iterative approaches that allow asymmetric CIs. As an alternative to analytic approaches, bootstrap methods can be used to generate empirical sampling distributions (Efron & Tibshirani, 1993; Rodgers, 1999). We include three bootstrap methods in the present study, two of which can provide intervals that are asymmetric about *A*.

Analytic CI Methods

Two methods that involve calculating an *SE* to then construct a CI are given by Hanley and McNeil (1982). Their work is grounded in the construction and analysis of ROC curves. Because they calculate *AUC* using the trapezoidal method, and therefore $A = AUC$, their formulas for the *SE* of *AUC* can be used to calculate the *SE* of *A*. There are two different formulas given by Hanley

[2] Two additional analytic methods were not included in our study because they deal with a slightly different statistic or a special type of data. First, a method introduced by Mee (1990) yields a CI centered on a variant of *A* that is calculated with no credit for tied scores. For example, using the illustrative data set presented earlier, whereas $A = 199/225 = .88$ when half credit is given for the 18 pairwise comparisons with tied scores, $A = 190/225 = .84$ when no credit is given for these ties. Because it is designed for use with a slightly different statistic, it would not be appropriate to evaluate the performance of this method in a study of CIs for *A*. Second, a method introduced by Ryu and Agresti (2008) provides an alternative for use with multinomial data.

and McNeil to calculate the *SE*.[3] One formula is nonparametric, and we call this the HM1 method. The other formula assumes population normality and equal variances; Hanley and McNeil refer to this as the bi-negative exponential formula, and we call this the HM2 method. Once the *SE* is calculated using either formula, it is inserted into a Wald-type expression to construct a 95% CI:

$$\mathrm{CI}_{.95} = A \pm 1.96 \times SE. \tag{3}$$

The SPSS module on ROC curves provides *AUC* as well as a CI constructed using either of the Hanley and McNeil formulas (nonparametric or bi-negative exponential).

A traditional formula to calculate an *SE* for *A* appears in many sources (e.g., Grissom & Kim, 2001, p. 141), and this can also be inserted into Equation 3 to construct a Wald-type CI. We call this the TR method, shorthand for "traditional."

Several other analytic approaches do not involve the calculation of an *SE* to construct a CI for *A*. One is presented by Fligner and Policello (1981), which we call the FP method, and another by Cliff (1993), which we call the CL method. These two approaches first calculate the δ statistic, an index for comparing two distributions (Cliff, 1993) that is related to *A* as follows: $A = (\delta + 1)/2$. The FP and CL methods construct CIs for δ, and the endpoints of these intervals can be converted back into the units of *A* using the equation shown earlier. Vargha and Delaney (2000) present an approach that they refer to as the Rank Welch method, which we call the RW method. Brunner and Munzel (2000) present an analytic approach to constructing CIs for *A*; we call this the BM method.

Whereas each of these seven analytic methods yields symmetric CIs that may extend beyond the theoretical range of values for *A*, the final two allow asymmetric CIs and respect the theoretical boundaries. Newcombe (2006a, 2006b) presented and studied a number of analytic methods, including many refinements of the Hanley and McNeil (1982) methods. Because the one that performed best in Newcombe's (2006b) simulation study was the fifth method listed, we follow others' lead in calling this the M5 method (Brown, Newcombe, & Zhao, 2009; Ryu & Agresti, 2008; Zhou, 2008). The M5 method uses an iterative technique to locate each end of the CI, which respects theoretical boundaries and allows asymmetry. Brown et al. introduced another iterative technique; we call this the BNZ method.

The top portion of Table 1 summarizes each of these nine analytic methods, including their assumptions as well as the results for the illustrative data set

[3] For the Hanley and McNeil formulas, as well as all other analytic methods, we refer readers to the primary sources cited here for details. We present only the details required to understand and evaluate our implementation of the bootstrap methods because this is the novel component of our investigation.

TABLE 1
Overview of Analytic and Bootstrap Methods

Analytic Method	*Assumptions*	*SE*	*95% CI*
Hanley and McNeil (1982) Nonparametric (HM1)	Normal sampling distribution No ties	.060	.767, 1.001
Hanley and McNeil (1982) Bi-negative exponential (HM2)	Normal sampling distribution No ties	.064	.759, 1.010
	Populations normal with equal variances		
Traditional (TR)	Normal sampling distribution No ties	.107	.674, 1.094
Fligner and Policello (1981; FP)	Normal sampling distribution		.774, .994
	Populations normal		
Cliff (1993; CL)	Normal sampling distribution Populations normal		.759, 1.010
Rank Welch (RW; Vargha & Delaney, 2000)	Sampling distribution follows *t* distribution No ties		.721, 1.048
Brunner and Munzel (2000; BM)	Sampling distribution follows *t* distribution		.764, 1.005
Newcombe (2006b; M5)	No ties Populations normal with equal variances		.694, .959
Brown, Newcombe, & Zhao (2009; BNZ)	No ties		.700, .949
	Populations normal with equal variances		

Bootstrap Method	*Assumptions*	*SE*	*95% CI*
Bootstrap *SE* (BSE)	Normal sampling distribution Sample representative of population	.058	.770, .998
Bootstrap percentile (BP)	Sample representative of population		.751, .978
Bootstrap bias-corrected and accelerated (BCA)	Sample representative of population		.709, .964

Note. *SE* = standard error; CI = confidence interval. For bootstrap methods, B = 1,999 bootstrap samples were used.

shown earlier. Figure 2 plots the CIs constructed using each method, which shows that seven of the nine analytic methods produced symmetric intervals, six of which extended above the maximum possible value of $A = 1$. This illustrates one of the common weaknesses of many analytic methods, the possibility that a CI can extend into impossible values. In addition, analytic methods require one

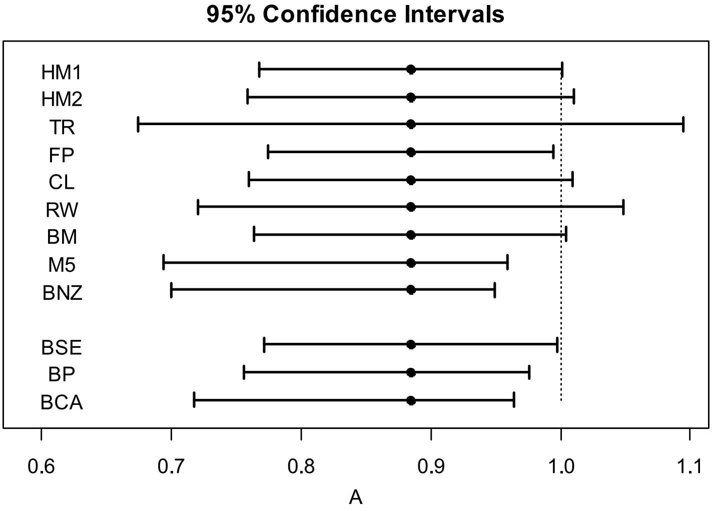

FIGURE 2 Confidence intervals constructed using all 12 methods included in this study, with the point estimate of $A = .884$ plotted for each; the dotted vertical line is plotted at the theoretical maximum value of $A = 1$. *Note.* HM1 = Hanley and McNeil (1982) nonparametric; HM2 = Hanley and McNeil (1982) bi-negative exponential; TR = traditional; FP = Fligner and Policello (1981); CL = Cliff (1993); RW = Rank Welch (Vargha & Delaney, 2000); BM = Brunner and Munzel (2000); M5 = Newcombe (2006b); Method 5; BNZ = Brown, Newcombe, & Zhao (2009); BSE = bootstrap standard error; BP = bootstrap percentile; BCA = bootstrap bias-corrected and accelerated.

or more assumptions that are frequently violated in practical applications (e.g., normal populations with equal variances, no tied scores, symmetric sampling distributions). These concerns motivated the search for alternative methods that make fewer or more realistic assumptions, that yield asymmetric and boundary-respecting CIs, and that might therefore provide better CI coverage.

Empirical CI Methods

Rather than making assumptions about the shape of a theoretical sampling distribution and estimating its parameters, bootstrap methods treat a sample of data as an unbiased estimate of the population (Efron & Tibshirani, 1993; Rodgers, 1999). A large number of samples is drawn from the observed data such that each is of the same size as the observed data; scores are sampled with replacement. Each bootstrap sample is submitted to analysis to contribute one statistical value to the empirical sampling distribution. For example, one can sample n_x scores from the observed distribution for one group and n_y scores from the observed distribution for the other group (in both instances sampling with replacement), calculate A for this two-group bootstrap sample of data, and then repeat this procedure a large number of times B to generate an empirical sampling distribution for A. This sampling distribution can be used in several ways to construct a CI for A. First, one can calculate the *SE* of A as the *SD* of all values in the empirical sampling distribution and substitute this into Equation 3 to construct a Wald-type CI. We call this the BSE (for bootstrap *SE*) method. Like most of the analytic methods, the BSE method provides symmetric CIs that may extend into impossible values.

In contrast, two other bootstrap methods can provide asymmetric CIs and do not yield CIs that extend into impossible values. In what is known as the percentile method, which we call the BP method, one sorts the B values in the empirical sampling distribution and identifies the values at the 2.5th and 97.5th percentiles as the limits of a 95% CI. Because the limits of the CI are located based on ordinal position, not a multiple of an *SE*, the lower and upper limits may or may not be equidistant from A. Finally, a bias-corrected and accelerated method described by Efron and Tibshirani (1993), which we call the BCA method, adjusts the percentiles used to form the limits of the CI based on factors such as the skewness of the empirical sampling distribution. Because the BCA method provides more accurate CIs than the percentile method for some applications, we included both in this study. The bottom portion of Table 1 gives CIs for each bootstrap method for the illustrative data set; $B =$ 1,999 bootstrap samples were used so that the tails of the empirical sampling distribution would be well defined and the thresholds for the 2.5th and 97.5th percentiles would fall between, not at, positions in the rank-ordered series of values. Because the BP and BCA bootstrap methods are based on locations within an empirical sampling distribution, their CIs cannot extend above the maximum possible value of $A = 1$. Figure 2 reveals the asymmetry of the BP and BCA intervals; the distance from A to the lower limit of the CI is greater than the distance to the upper limit.

THE PRESENT STUDY

A number of studies (e.g., Brown et al., 2009; Newcombe, 2006b; Ryu & Agresti, 2008; Zhou, 2008) have examined the performance of methods for constructing CIs for A, including many of those listed earlier. The present study was designed to build on previous research by including a broader range of CI methods and spanning a broader range of data conditions. We included nine analytic methods that have either shown promising results in prior investigations or that have not yet been studied rigorously as well as three empirical methods, few of which have been included in prior work. Our simulation study included data conditions selected to pose realistic challenges for all 12 methods. The study was designed such that assumptions underlying various methods were sometimes satisfied and sometimes not, and violations of assumptions were substantial but not extreme. To identify the CI methods that perform best, key criteria involved the extent to which observed CI coverage probabilities approximated the nominal level of 95% as robustly as possible across data conditions. Secondary criteria involved the length of the CIs (shorter CIs are preferable), the extent to which coverage errors were evenly distributed below the lower limits and above the upper limits, and how often CIs extended into impossible values.

METHOD

Design and Data Generation

A simulation study was performed using programs written for the R computing environment (R Development Core Team, 2011). There were 324 cells in the fully crossed factorial design, which contained six factors whose levels spanned challenging yet realistic data conditions. We included data conditions that were both favorable (e.g., normal populations, equal variances, no tied scores) and unfavorable (e.g., skewed populations, unequal variances, tied scores) to the methods under study while avoiding conditions sufficiently extreme that they might "stack the deck" against any particular methods. Moreover, to ensure the ability to perform the study in a timely manner as well as to analyze and present the results coherently, we limited the number of factors and levels to those that seemed most informative.

1. **Effect size.** The degree of separation between groups was indexed using Cohen's d, with three levels of $d = 0.00$, 0.50, and 2.00. These levels correspond to no effect, a medium effect (by the conventional rules of thumb; Cohen, 1988), and a very large effect that might pose a challenge for some techniques (i.e., because the actual sampling distribution of A will be negatively skewed for large effects, symmetric CIs may extend above the theoretical maximum value of 1.00).
2. **Sample size.** Total sample size spanned small to moderate values, specifically $N = 30$, 60, or 120.
3. **Group sizes.** Groups were either equal in size or unequal such that one group contained 3 times as many cases as the other. In other words, the base rate of the larger group was either $P = .25$, .50, or .75. For example, with a total N of 60, group sizes were 15/45, 30/30, or 45/15.
4. **Variance ratio.** Populations were created with variances that were either equal ($VR = 1{:}1$) or unequal ($VR = 4{:}1$). When $VR = 4{:}1$, the higher scoring population possessed greater variance. Because all factors in the

design were fully crossed, this means that the design included cases of positive as well as negative relationships between variance ratios and relative group sizes.

5. **Distributions.** Populations were created such that both were normal, both were positively skewed, or one was skewed in each direction. Skewed distributions were generated using the *g*-and-*h* transformations shown in Hoaglin (1985); briefly, *g* controls asymmetry and *h* controls tail weight relative to a normal distribution (in which $g = .00$ and $h = .00$). For positive skew, we used $g = .30$ and $h = .00$ (which corresponds to skewness of .95 and kurtosis of 1.64); for negative skew, we used $g = -.30$ and $h = .00$.
6. **Response scales.** Scores were either left in truly continuous form or cut into seven ordered categories using equally spaced thresholds applied to the distribution of scores pooled across populations. Using ordered categories represents the fact that in actual research data (e.g., collected using Likert-type response scales) there are often a nontrivial number of tied scores. This might pose a challenge for CI techniques that assume there are no tied scores.

Within each of the 3 (effect size) × 3 (sample size) × 3 (group sizes) × 2 (variance ratio) × 3 (distributions) × 2 (response scales) = 324 cells in the study's design, a pair of finite populations each with $N = 100{,}000$ was created using equally spaced quantiles, and 1,000 replication samples were drawn at random for analysis. To implement bootstrap methods, $B = 1,999$ bootstrap samples were drawn for each replication sample.

Data Analysis

For each of the 1,000 replication samples, *A* was calculated using Equation 1. For each value of *A,* nine different analytic methods were used to construct CIs. Of these, three methods (HM1, HM2, and TR) involved calculating *SE*s to construct CIs using Equation 3 and six methods (FP, CL, RW, BM, M5, and BNZ) did not. The empirical approach was used to construct three more CIs. The empirical sampling distribution of $B = 1{,}999$ values of *A* for each replication sample was used to construct CIs for *A* following the BSE, BP, and BCA methods.

RESULTS

Several measures were calculated for each of the 12 CI methods studied. First, the mean coverage level was calculated for each method as the percentage of the samples' CIs that contained the population value of *A*; coverage in the 94%

TABLE 2
Summary of Confidence Interval Results for All Data Conditions

Method	*Mean %* *Coverage*	*% Within* *Control Limits*	*Mean %* *< LL*	*Mean %* *> UL*	*Mean* *Length*	*% UL* *> 1*	*Mean* *Symmetry*
HM1	93.45	26.85	5.01	1.53	.2810	12.26	1.00
HM2	93.75	25.31	4.48	1.77	.2781	13.26	1.00
TR	97.66	20.99	1.06	1.28	.3391	26.36	1.00
FP	92.46	32.72	5.23	2.31	.2738	13.43	1.00
CL	92.93	59.57	5.25	1.81	.2729	12.84	1.00
RW	96.20	54.01	1.99	1.81	.2950	20.67	1.00
BM	92.80	59.57	4.90	1.81	.2782	13.14	1.00
M5	96.68	28.70	1.19	2.13	.2722	0.00	0.77
BNZ	92.40	37.65	2.11	5.50	.2355	0.00	0.79
BSE	91.42	32.41	6.17	2.41	.2512	10.65	1.00
BP	92.49	46.60	5.05	2.46	.2492	0.00	0.90
BCA	94.40	68.83	2.92	2.67	.2592	0.00	0.81

Note. LL = lower limit of confidence interval; UL = upper limit of confidence interval; HM1 = Hanley and McNeil (1982) nonparametric; HM2 = Hanley and McNeil (1982) bi-negative exponential; TR = traditional; FP = Fligner and Policello (1981); CL = Cliff (1993); RW = Rank Welch (Vargha & Delaney, 2000); BM = Brunner and Munzel (2000); M5 = Newcombe (2006b), method 5; BNZ = Brown, Newcombe, & Zhao (2009); BSE = bootstrap standard error; BP = bootstrap percentile; BCA = bootstrap bias-corrected and accelerated.

to 96% range was considered an excellent approximation to the nominal 95% level. Second, the percentage of all cells for which each method's CI coverage was within 95% control limits[4] was calculated to assess robustness across data conditions; the higher the percentage within control limits, the better. Third and fourth, the percentages of samples for which the population value of *A* fell below the lower bound of the CI or above the upper bound of the CI were calculated; percentages in the 2% to 3% range were considered excellent approximations to the nominal 2.5% coverage error rate at each end. Fifth, the mean length of all CIs was calculated; the shorter the CIs, the better. Sixth, percentages of samples for which CIs extended above the theoretical maximum value of $A = 1$ was calculated; the smaller the percentage, the better. Seventh, the mean ratio of the upper to lower CI segments' lengths was calculated as an index of the symmetry of the CIs; values less than 1.00 reflect sensitivity to the asymmetry of the sampling distribution, though we are aware of no criterion for the "best" sensitivity. Results across all data conditions are summarized in Table 2.

[4] With 1,000 replication samples per cell, 95% control limits correspond to CI coverage values of 93.65% to 96.35%.

By nearly all measures, the BCA method provided the most accurate CIs. Population values fell within the CIs of the BCA method 94.40% of the time, which was closer to the nominal coverage value of 95% than for any other method. Coverage ranged as low as 91.42% (for the BSE method) to as high as 97.66% (for the TR method), but only the BCA method yielded an overall coverage value between 94% and 96%. Across conditions, the BCA method's coverage was within the 95% control limits 68.83% of the time; values for other methods ranged from 20.99% (for the TR method) to 59.57% (for the CL and BM methods). The BCA method was the only approach for which coverage error rates fell between 2% and 3% at both ends of the CIs (2.92% for lower limits, 2.67% for upper limits), and no other method yielded rates that were proportionally as similar to one another (i.e., 2.92/2.67 = 1.09, less than any other method's ratio of the larger to smaller rate). The BCA method produced CIs that were shorter (mean length = .2592) than those for all but one analytic method, which ranged from .2722 (for the M5 method) to .3391 (for the TR method). Shorter CIs were produced by the BNZ method (.2355) and the other bootstrap methods (.2492 for BP and .2512 for BSE), but these alternatives did not fare as well as the BCA method by any other measures; their shorter lengths corresponded to liberal coverage probabilities.

Four methods respect the theoretical boundaries for A (0 to 1), but among the other methods the upper limits of CIs frequently extended above the theoretical maximum of 1. This occurred almost exclusively for the one third of data conditions with $d = 2.00$, in which the percentages shown in Table 2 were nearly triple their size. The extent of CI asymmetry was measured as the ratio of the distance from A to the upper limit to the distance from A to the lower limit. Ratios less than one are indicative of CIs shorter on the upper end than the lower end, which corresponds to a negatively skewed sampling distribution. Across all data conditions, the BP method yielded less asymmetry (0.90) than the BCA method (0.81), the BNZ method (0.79), or the M5 method (0.77). These means masked significant heterogeneity across population effect sizes. For data conditions with $d = 2.00$, when sampling distributions were negatively skewed, all four methods yielded intervals shorter on the upper than lower end (.75 for BP, .54 for BCA, .43 for M5, and .37 for BNZ). For data conditions with $d = 0.00$, when sampling distributions should be fairly symmetric, CIs were symmetric for the BP, BCA, and M5 methods (1.00, 1.00, and 1.03, respectively) but longer on the upper than lower end for the BNZ method (1.25). Because there is no criterion for evaluating performance by this measure, it is unclear whether the BNZ method's substantially greater sensitivity to any departures from normality in the sampling distribution, including chance-level deviations from normality among samples drawn from a population with a null effect size, represents a strength or a liability.

In addition to evaluating CI construction across all data conditions, we also examined results within each level of each design factor in the study as well as when all four assumptions made by various analytic methods were satisfied (normal populations, equal variances, no tied scores, and normal sampling distributions). This provided 17 separate comparisons of all 12 CI methods. To help assess robustness across these data conditions, we calculated the percentage of cells for which coverage was within 95% control limits. These results are shown in Table 3, with bold print highlighting the best performing method within each comparison—as well as methods that did not differ statistically significantly from the best (by z test for dependent proportions, two-tailed $\alpha = .05$)—and italics highlighting especially poorly performing methods (less than 50% as

TABLE 3
Percentage of Cells With Coverage Within Control Limits

	Analytic									*Empirical*		
Factors and Levels	*HM1*	*HM2*	*TR*	*FP*	*CL*	*RW*	*BM*	*M5*	*BNZ*	*BSE*	*BP*	*BCA*
Sample sizes												
$N = 30$	*13.0*	*19.4*	*20.4*	*18.5*	**42.6**	**47.2**	**41.7**	26.9	24.1	*10.2*	*22.2*	**43.5**
$N = 60$	*28.7*	*29.6*	*20.4*	*34.3*	63.0	57.4	60.2	*33.3*	43.5	*25.9*	47.2	**75.9**
$N = 120$	*38.9*	*26.9*	*22.2*	*45.5*	73.1	57.4	76.9	*25.9*	45.4	61.1	70.4	**87.0**
Effect sizes												
$d = 0.00$[a]	*33.3*	*28.7*	*36.1*	50.0	**76.9**	**76.9**	**83.3**	*36.1*	*37.0*	44.4	63.9	73.1
$d = 0.50$	*31.5*	*29.6*	*26.9*	41.7	70.6	**81.5**	**75.0**	*34.3*	*28.7*	*37.0*	52.8	70.4
$d = 2.00$	*15.7*	*17.6*	*0.0*	*6.5*	*22.2*	*3.7*	*20.4*	*15.7*	47.2	*15.7*	*23.1*	**63.0**
Group sizes												
$P = .25$	*25.9*	*28.7*	*16.7*	*27.8*	51.9	**61.1**	**56.5**	*26.9*	*29.6*	*19.4*	31.5	**63.0**
$P = .50$	*28.8*	*27.8*	*30.6*	38.9	**70.4**	55.6	**69.4**	38.9	41.7	45.4	62.0	**74.1**
$P = .75$	*25.9*	*19.4*	*15.7*	*31.5*	56.5	45.4	52.8	*20.4*	41.7	*32.4*	46.3	**69.4**
Variance ratios												
$VR = 1:1$[b]	*32.1*	*32.1*	*28.4*	*34.0*	60.5	56.8	57.4	37.7	47.5	*31.5*	44.4	**68.5**
$VR = 4:1$	*21.6*	*18.5*	*13.6*	*31.5*	58.6	51.2	61.7	*19.8*	*27.8*	*33.3*	48.8	**69.1**
Distributions												
Normal[c]	*34.3*	*30.6*	*26.9*	41.7	54.6	56.5	54.6	39.8	43.5	*24.1*	43.5	**76.9**
+Skewed	*33.3*	*27.8*	*17.6*	34.3	58.3	59.3	61.1	*23.1*	38.0	*31.5*	46.3	**68.5**
±Skewed	*13.0*	*17.6*	*18.5*	*22.2*	**65.7**	46.3	**63.0**	*23.1*	*31.5*	41.7	50.0	**61.1**
Scales												
Continuous[d]	*27.2*	*27.2*	*36.4*	*33.3*	53.7	53.7	51.2	48.8	60.5	*25.9*	42.0	**73.5**
Categorical	*26.5*	*23.5*	*5.6*	*32.1*	**65.4**	54.3	**67.9**	*8.6*	*14.8*	38.9	51.2	**64.2**
All assumptions satisfied[a,b,c,d]	*33.3*	*33.3*	**100.0**	55.6	77.8	**100.0**	66.7	**100.0**	**100.0**	*33.3*	66.7	**100.0**
Top	0	0	1	0	5	5	7	1	1	0	0	15
Bottom	17	17	16	11	1	1	1	11	6	12	1	0

Note. Bold print highlights the highest percentage for a row or a value not statistically significantly different from the highest (using z tests for dependent proportions, $\alpha = .05$, two-tailed). Italics highlight percentages less than 50% of the highest value for a row. The numbers of bold and italicized entries are tallied for each method (column) and presented as the "top" and "bottom" sums, respectively, at the bottom of the table. HM1 = Hanley and McNeil (1982) nonparametric; HM2 = Hanley and McNeil (1982) bi-negative exponential; TR = traditional; FP = Fligner and Policello (1981); CL = Cliff (1993); RW = Rank Welch (Vargha & Delaney, 2000); BM = Brunner and Munzel (2000); M5 = Newcombe (2006b), method 5; BNZ = Brown, Newcombe, & Zhao (2009); BSE = bootstrap standard error; BP = bootstrap percentile; BCA = bootstrap bias-corrected and accelerated.

[a] Symmetric (normal or t) sampling distribution. [b] Equal population variances. [c] Normal population distributions. [d] No tied scores.

TABLE 4
Summary of Standard Error Results for All Data Conditions

Method	*Mean SE*	*Bias*	*Accuracy*
HM1	.0717	.0057	.0082
HM2	.0710	.0050	.0102
TR	.0865	.0205	.0229
BSE	.0641	−.0019	.0027

Note. HM1 = Hanley and McNeil (1982) nonparametric; HM2 = Hanley and McNeil (1982) bi-negative exponential; TR = traditional; BSE = bootstrap standard error.

good as the best). The BCA method was the best or among the best in 15 of 17 comparisons, and it was never among the especially poor performers. No other method was nearly as robust. Each was among the especially poor performers at least once and among the best less than half the time.[5]

For the four methods that involve the calculation of *SE*s (HM1, HM2, TR, and BSE), pertinent results are shown in Table 4. The BSE method yielded the smallest mean *SE* (.0641), followed by comparable values for the HM1 and HM2 methods (.0717 and .0710, respectively), and then the TR method (.0865). As a point of comparison, the observed *SE* was calculated within each cell of the design as the *SD* of the *A* values for the 1,000 replication samples; the mean of these observed *SE* values was .0660. Residual *SE*s were calculated as the *SE* calculated using one method minus the observed *SE* for that cell. Bias was then calculated as the mean of these residuals, and accuracy was calculated as the mean of the absolute values of these residuals. The HM1 and HM2 methods mildly overestimated *SE*s (by 8% to 9% of their observed values, respectively) and were fairly accurate

[5] As an additional examination of robustness to more extreme data conditions and at the suggestion of an anonymous reviewer, we ran supplementary analyses with four ordered categories (rather than continuous response scales or seven ordered categories, as in the main study). Crossed with all other factors in the design, this yielded results for 162 new cells. Because these analyses were supplementary, we included only 100 replication samples per cell (rather than 1,000 as in the main study). The results showed that the BCA method remained competitive, and arguably the best choice, under these new conditions. Only five methods attained mean coverage levels between 94% and 96%: CL, RW, BM, BP, and BCA; these were also the five methods at or not significantly different from the best performer in terms of the percentage of cells within 95% control limits. Among these five methods, the BP and BCA methods yielded shorter CIs (mean length = .2352 and .2433, respectively) than the CL, RW, and BM methods (mean length = .2579, .2673, .2648, respectively). Whereas the bootstrap methods respected the theoretical boundaries of the *A* statistic, the CL, RW, and BM methods produced CIs that extended above 1.00 for 3.70%, 4.98%, and 4.46% of all samples. Thus, even with very few ordered categories, there appears to be no reason to prefer an alternative to the BCA method.

(average error of 12% and 15%, respectively). The TR method overestimated by a larger amount (31%) and was less accurate (average error of 35%). Though the BSE method underestimated by a small amount (3%), it was the least biased and the most accurate (average error of 4%) method for calculating an *SE*.

AN EMPIRICAL ILLUSTRATION

Hirsch (2005) introduced a citation-based measure of scholarly impact designed to reward both the quantity and quality of an individual's published papers. This *h* index is calculated as the largest number *h* such that an author has published at least *h* papers that have been cited at least *h* times each. Publishing a large number of rarely cited papers will not yield a large score on the *h* index; nor will publishing a small number of highly cited papers. A high score on the *h* index can be attained only by publishing many influential papers. How useful a measure of scholarly impact is the *h* index? Ruscio, Seaman, D'Oriano, Stremlo, and Mahalchik (2011) collected several large samples of citation data to assess the *h* index and many alternative indices of scholarly impact in a variety of ways. Their largest sample included citation counts for 10 randomly selected professors from each of 175 universities' psychology departments. The departments were ranked by the National Research Council (Goldberger, Maher, & Flattau, 1995). For present purposes, a small subset of their data was analyzed. Specifically, all of the full professors sampled from the top 11 and bottom 11 programs were selected, which yielded $n = 45$ full professors at the high-ranked universities and $n = 47$ full professors at the low-ranked universities. Scores on the *h* index were compared across these two groups.

The top graph in Figure 3 shows the score distributions for each group, which ranged from 1 to 53 ($Mdn = 17$, $IQR = 13$ to 26) for professors at high-ranked universities and from 0 to 14 ($Mdn = 3$, $IQR = 2$ to 6) for professors at low-ranked universities. For these data, $A = .895$, which corresponds to an 89.5% chance that a randomly selected full professor from a high-ranked university would score higher on the *h* index than a randomly selected full professor from a low-ranked university. This is a very large effect, and it would be useful to know something about how precisely the population effect size has been estimated from these data. Constructing a CI would address this issue. Unfortunately, most of the assumptions required by the analytic CI methods were violated.

Because the *h* index can only take integer values, there were many tied scores; out of the $45 \times 47 = 2{,}115$ pairwise score comparisons, there were 54 ties. Both groups' distributions were nonnormal (skewness = 0.71 and 1.13 for professors at high- and low-ranked universities) and the variances were heterogeneous (*SD*s = 12.26 and 3.81). Due to the large estimated effect size, it would be unreasonable to assume a symmetric sampling distribution for *A*.

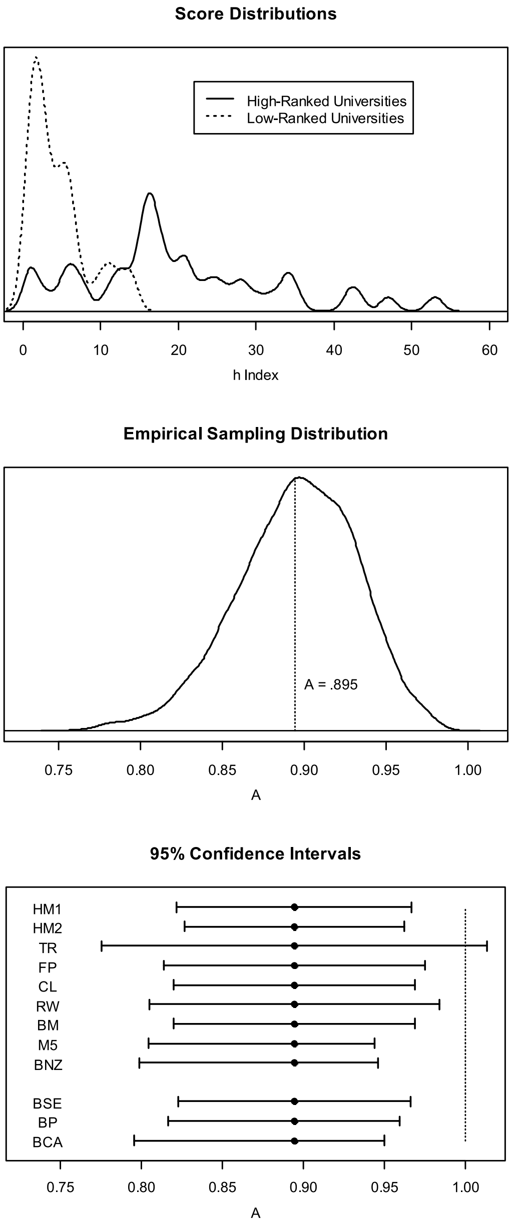

FIGURE 3 (See Figure 3 caption on page 219.)

As expected, when B = 1,999 bootstrap samples were used to construct an empirical sampling distribution, it was negatively skewed (skewness = –0.39; see Figure 3, middle graph). The only CI method likely to be robust in the face of these challenging—but by no means uncommon—conditions is the BCA method, which assumes only that the sample is representative of the population. Because these individuals were selected at random from the target population, there is little or no reason to doubt the representativeness of the sample.[6]

CIs constructed using all 12 methods are shown in the bottom graph of Figure 3. Perhaps the most striking feature of this graph is the variability across the intervals. Their symmetry (or asymmetry) and relative lengths reflect the findings of the simulation study. There is no question that the choice of a CI method has consequences, that one is not splitting hairs when asking which should be preferred. The TR method yielded an interval so wide that its upper limit surpassed the maximum possible value of $A = 1$. All of the other intervals spanned admissible values, but only the M5, BNZ, BP, and BCA methods produced asymmetric intervals that reflect the asymmetry of the sampling distribution. A single sample of data does not afford a conclusion as to which of these CIs is the most appropriate; we present these results only to illustrate the trends observed in our simulation study, not as a follow-up test. Because the simulation results suggest that the BCA method is most likely to provide good coverage, we would recommend using the CI produced using this method.

[6] One might argue that the sample may be unrepresentative due to the luck of the draw, that with $N = 92$ it is not possible to evaluate the potential sample bias that could emerge by chance. However, it is important to consider that there is a finite, and in fact a rather small, population of full professors in the psychology departments of the top 11 and bottom 11 universities. For example, if a typical university's psychology department has a faculty of 30 members and approximately half are full professors, one would expect there to be a population of about 330 full professors at 22 universities. In the context of this admittedly rough estimate, a sample of 92 is not so small and is likely to be reasonably representative.

FIGURE 3 (See Figure 3 artwork on page 218.) Analysis of scores on the h index for 45 full professors at high-ranked universities and 47 full professors at low-ranked universities. The top graph shows the score distributions (densities) for each group. The middle graph shows the empirical sampling distribution (density) of A obtained using B = 1,999 bootstrap samples. The bottom graph shows confidence intervals constructed using all 12 methods included in this study, with the point estimate of $A = .895$ plotted for each; the dotted vertical line is plotted at the theoretical maximum value of $A = 1$. *Note.* HM1 = Hanley and McNeil (1982) nonparametric; HM2 = Hanley and McNeil (1982) bi-negative exponential; TR = traditional; FP = Fligner and Policello (1981); CL = Cliff (1993); RW = Rank Welch (Vargha & Delaney, 2000); BM = Brunner and Munzel (2000); M5 = Newcombe (2006b); Method 5; BNZ = Brown, Newcombe, & Zhao (2009); BSE = bootstrap standard error; BP = bootstrap percentile; BCA = bootstrap bias-corrected and accelerated.

DISCUSSION

The goal of this investigation was to determine the best methods for constructing CIs for the effect size estimator *A*, which equals the area under an ROC curve. Results support the BCA method, which yielded a mean coverage value closer to the nominal 95% level than any other method, the most coverage values within 95% control limits, coverage error rates close to 2.5% at both ends of the CIs, and intervals that were shorter than those produced by most other methods. In addition, the BCA method is among those that respect the theoretical boundaries for *A* and allow asymmetric CIs. Compared with the 11 other methods evaluated in this study, the performance of the BCA method was the most robust across a broad array of data conditions, including those that violated one or more of the assumptions made by many analytic methods (e.g., normal populations, equal variances, no tied scores, symmetric sampling distributions). Along with several other methods, the BCA method performed exceptionally well when these four assumptions were satisfied. In the full series of 17 comparisons within specific data conditions, the BCA method was usually (15 times) among the best performers and was never among the especially poor performers. Programs to calculate *A*, calculate its *SE* using the bootstrap, and construct a CI using the BCA method are available at http://www.tcnj.edu/~ruscio/taxometrics.html

The two conditions under which the BCA method was not among the best performers suggest one important cautionary note. Other methods' coverage rivaled or surpassed that of the BCA method when the null hypothesis was true (in this study, when $d = 0.00$) or nearly true (when $d = 0.50$). Thus, constructing CIs using the BCA method may not be the best way to test the null hypothesis of $\Delta = .50$, which represents stochastic equality (i.e., $\Pr(X > Y) = \Pr(Y > X)$). Even when applied to the same data, different data-analytic methods may be most appropriate for different research purposes. For example, in the context of correlation analysis, Lee and Rodgers (1998) recommended different bootstrap methods for CI construction than for testing the null hypothesis of $\rho = 0$. Future research would be required to determine whether any of the CI methods studied here is preferable to methods designed explicitly to test the null hypothesis of $\Delta = .50$ (see Brunner & Munzel, 2000; Cliff, 1996; Delaney & Vargha, 2002; Fagerland & Sandvik, 2009; Fligner & Policello, 1981; Neuhauser, Losch, & Jockel, 2007; Ryu & Agresti, 2008; for related work on sample size determination, see Vollandt & Horn, 1997). We recommend choosing a method that appears best suited to the purpose at hand. The present study suggests that if one wants to calculate an *SE* for *A* (e.g., to weight effect size estimates in a meta-analysis) or construct a CI for *A*, bootstrap methods seem to be a good choice; specifically, one can use the BSE method to calculate an *SE* and the BCA method to construct a CI. This study sheds no light on the selection of a method to test the null hypothesis of $\Delta = .50$.

This would require, at minimum, the examination of Type I error rates for null effects and statistical power for nonnull effects (see Delaney & Vargha, 2002, for an excellent overview of available methods).

The robust performance of the BCA method for CI construction appears to stem from two features. First, the empirical approach of this bootstrap technique frees the user from the more restrictive assumptions of the analytic methods. Bootstrapping accommodates tied scores, affords robustness to population distributions that deviate from normality and equal variances, respects the theoretical boundaries of the A statistic, and enables the construction of asymmetric CIs when sampling distributions are skewed. Second, the bias correction and acceleration of the BCA method provided helpful adjustments to the CI limits relative to those obtained using the BP method. Both of these bootstrap methods share all of the desirable characteristics listed earlier, yet the BCA method performed considerably better than the BP method. Because both methods locate CI limits within the same empirical sampling distribution, the improved performance of the latter must be due to its adjustment of the CI limits through bias correction and acceleration. Bootstrapping is more computationally intensive than the analytic methods, but not prohibitively so. Analyzing a sample of data with $N = 120$ using $B = 1{,}999$ bootstrap samples takes < 1 s on a laptop computer running the programs cited earlier.

One well-known weakness of the bootstrap, including the BCA method, is its sensitivity to sample size. The assumption that the sample is representative of the population is increasingly untenable with smaller samples. In this study, total samples were small to modest in size ($N = 30$, 60, or 120). With $N = 60$ or 120 the BCA method outperformed all others. With $N = 30$ its performance was not statistically significantly different from that of the RW method, which was the top performer, or the CL and BM methods, which performed only slightly (and not significantly) more poorly than the BCA method. With even smaller samples, it is possible that the RW method or others could significantly surpass the BCA method. On the other hand, even though we did not study large samples ($N > 120$), it seems safe to assume that the BCA method will continue to perform well because these will be even more representative of the populations from which they are drawn. For samples of at least modest size ($N \geq 60$), investigators calculating A as an effect size estimate or AUC as a component of an ROC analysis can rely on the BCA method to construct CIs with good coverage. With samples as small as $N = 30$, the BCA still performed about as well as or better than all other methods tested here and seems to be a safe choice.

This study's design spanned a broad array of data conditions, including many that violated one or more of the assumptions of analytic methods. Because bootstrap methods make fewer assumptions for constructing CIs for A, one might expect that the BCA method would continue to outperform others under more severe violations (e.g., more skew, more unequal variances, more tied scores). In addition to further study under more extreme data conditions, it might be worthwhile to examine CI methods for applications of the *A* statistic in other research designs. McGraw and Wong (1992) and Vargha and Delaney (2000) introduced ways to use the *CL* and *A* statistics, respectively, with correlated rather than independent groups as well as with more than two groups. We dealt exclusively with two independent groups in the present research because virtually all prior research has done the same and this design appears to be the most common application in practice. Future research should explore the construction of CIs for other applications of the *A* statistic.

REFERENCES

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Brown, B. M., Newcombe, R. G., & Zhao, Y. (2009). Non-null semi-parametric inference for the Mann-Whitney measure. *Journal of Nonparametric Statistics, 21,* 743–755.

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and small-sample approximation. *Biometrical Journal, 42,* 17–25.

Cliff, N. (1993). Dominance statistics: Ordinal methods for behavioral data analysis. *Psychological Bulletin, 114,* 494–509.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sample sizes. *Psychological Methods, 7,* 485–503.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. San Francisco, CA: Chapman & Hall.

Fagerland, M. W., & Sandvik, L. (2009). The Wilcoxon-Mann-Whitney test under scrutiny. *Statistics in Medicine, 28,* 1487–1497.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27,* 861–874.

Fligner, M. A., & Policello, G. E., II. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association, 76,* 162–168.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Goldberger, M. L., Maher, B. A., & Flattau, P. E. (Eds.). (1995). *Research doctorate programs in the United States: Continuity and change*. Washington, DC: National Academies Press.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135–146.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Hedges, L. V. (1981). Distributional theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102,* 16569–16572.

Hoaglin, D. C. (1985). Summarizing shape numerically: The *g*-and-*h* distributions. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data, tables, trends, and shapes* (pp. 461–509). New York, NY: Wiley.

Hsu, L. M. (2004). Biases of success rate differences shown in binomial effect size displays. *Psychological Methods, 9,* 183–197.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Lee, W.-C., & Rodgers, J. L. (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods, 3,* 91–103.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods, 11,* 386–401.

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

Mee, R. W. (1990). Confidence intervals for probabilities and tolerance regions based on a generalization of the Mann-Whitney statistic. *Journal of the American Statistical Association, 85,* 793–800.

Neuhauser, M., Losch, C., & Jockel, K.-H. (2007). The Chen-Luo test in case of heteroscedasticity. *Computational Statistics and Data Analysis, 51,* 5055–5060.

Newcombe, R. G. (2006a). Confidence intervals for an effect size measure based on the Mann-Whitney statistic: Part 1. General issues and tail-based area methods. *Statistics in Medicine, 25,* 543–557.

Newcombe, R. G. (2006b). Confidence intervals for an effect size measure based on the Mann-Whitney statistic: Part 2. Asymptotic methods and evaluation. *Statistics in Medicine, 25,* 559–573.

R Development Core Team. (2011). R: A language and environment for statistical computing [Software]. Retrieved from http://www.R-project.org

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34,* 441–456.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13,* 19–30.

Ruscio, J., Seaman, F., D'Oriano, C., Stremlo, E., & Mahalchik, K. (2011). *Evaluating scholarly impact: New tools to address old problems.* Manuscript submitted for publication.

Ryu, E., & Agresti, A. (2008). Modeling and inference for an effect size measure. *Statistics in Medicine, 27,* 1703–1717.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285–1293.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect size. *Educational Researcher, 31,* 25–32.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132.

Vollandt, R., & Horn, M. (1997). Evaluation of Noether's method of sample size determination for the Wilcoxon-Mann-Whitney test. *Biometrical Journal, 39,* 823–829.

Wilkinson, L., and the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wolfe, D. A., & Hogg, R. V. (1971). On constructing statistics and reporting data. *The American Statistician, 25,* 27–30.

Zhou, W. (2008). Statistical inference for $P(X < Y)$. *Statistics in Medicine, 27,* 257–279.