Multivariate Behavioral Research, 48:208–219, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2012.738184

Generalizations and Extensions of the Probability of Superiority Effect Size Estimator

John Ruscio and Benjamin Lee Gera
The College of New Jersey

Researchers are strongly encouraged to accompany the results of statistical tests with appropriate estimates of effect size. For 2-group comparisons, a probability-based effect size estimator (A) has many appealing properties (e.g., it is easy to understand, robust to violations of parametric assumptions, insensitive to outliers). We review generalizations of the A statistic to extend its use to applications with discrete data, with weighted data, with $k > 2$ groups, and with correlated samples. These generalizations are illustrated through reanalyses of data from published studies on sex differences in the acceptance of hypothetical offers of casual sex and in scores on a measure of economic enlightenment, on age differences in reported levels of Authentic Pride, and in differences between the numbers of promises made and kept in romantic relationships. Drawing from research on the construction of confidence intervals for the A statistic, we recommend a bootstrap method that can be used for each generalization of A. We provide a suite of programs that should make it easy to use the A statistic and accompany it with a confidence interval in a wide variety of research contexts.

An estimator of effect size indicates the magnitude of the relationship among variables. The inclusion of effect size statistics in research reports can help readers in several ways, including distinguishing between practical and statistical significance (Kirk, 1996). As a result, the American Psychological Association (APA) recommends that researchers accompany tests of statistical significance with appropriately chosen effect size estimates (APA, 2009).

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

The comparison of scores across members of two groups is common in psychological studies, and several effect size estimators are used by researchers. These include Cohen's *d*, the point-biserial correlation (r_{pb}), and a probability-based statistic that can be calculated in a parametric or nonparametric manner. Ruscio (2008) discusses the advantages and disadvantages of these estimators as well as ways to convert between them (given certain assumptions). Cohen's *d* is the standardized mean difference, calculated as the difference between two groups' mean scores on the dependent variable divided by the within-group standard deviation. This statistic is relatively easy to calculate and understand, but it is sensitive to violations of the normality or equal variances assumptions, unequal group sizes, and outliers. The point-biserial correlation r_{pb} is calculated as the correlation between group membership (coded using any two distinct numerical values) and scores on the dependent variable. This measures how well one variable predicts another and connects more directly to concepts such as statistical power and the general linear model, although it is harder to interpret, less relevant to understanding treatment effects, and sensitive to the same factors that can influence *d*.

In this research, we focus on a probability-based effect size statistic that compares favorably with *d* or r_{pb} in many ways. McGraw and Wong (1992) introduced the common-language effect size statistic *CL* as the probability that a randomly selected member of one group scores higher than a random selected member of another group. They calculated *CL* by making the usual parametric assumptions of normality and equal variances. Subsequently, many researchers have calculated a nonparametric variant that has been identified by various names (Delaney & Vargha, 2002; Grissom, 1994; Grissom & Kim, 2001; Hsu, 2004; Vargha & Delaney, 2000). Ruscio (2008) reviewed the similarities among these variations and followed Vargha and Delaney's (2000) lead by using the label *A* to emphasize this statistic's relationship to the area under a receiver operating characteristic curve. Specifically, *A* is calculated as

$$A = [\#(Y_1 > Y_2) + .5\#(Y_1 = Y_2)]/n_1 n_2, \qquad (1)$$

where # is the count function, Y_1 and Y_2 refer to scores by members of Groups 1 and 2, and n_1 and n_2 are the group sizes. In other words, one makes all pairwise comparisons between members of Group 1 and members of Group 2, tallies the number of times that the former scores higher than the latter (or credits this as .5 if they are tied), and divides by the total number of comparisons. Values can range from .00 (all members of Group 2 score higher than all members of Group 1) through .50 (equal probability that members of either group score higher, more formally known as *stochastic equality*) to 1.00 (all members of Group 1 score higher than all members of Group 2). *A* serves as an estimator of $\Delta = \Pr(Y_1 > Y_2)$, which Grissom and Kim (2012) called the "probability of superiority" (p. 149), and its calculation is simple and intuitive.

For example, consider a comparison between the following two groups of scores: $Y_1 = \{2, 3, 4\}$ and $Y_2 = \{1, 2, 3\}$. The first member of Y_1 outscores one member of Y_2 and ties another, yielding 1.5 points toward the numerator of A. The second member of Y_1 outscores two members of Y_2 and ties another, yielding 2.5 points toward the numerator of A. The third member of Y_1 outscores all three members of Y_2, yielding 3 points toward the numerator of A. Summing across all comparisons yields $A = (1.5+2.5+3)/(3\times3) = 7/9 = .78$. Allowing partial credit for ties, this means that there is a 78% chance that a randomly chosen score from Y_1 would exceed a randomly chosen score from Y_2.

This probability-based statistic does not offer the same connectivity to other statistical concepts as r_{pb}, but it does possess a number of advantages relative to d and r_{pb}. The A statistic does not require parametric assumptions, is highly robust to the influence of outliers, and is insensitive to unequal group sizes—which means that it can be more helpful when generalizing findings to other research contexts. Perhaps most important is that A is easier to understand than d or r_{pb}, facilitating communication even with those untutored in statistics. For example, McGraw and Wong (1992) cited data on the sex difference in height among U.S. adults. Using these data, the standardized mean difference is $d = 2.00$ and the point-biserial correlation is $r_{pb} = .71$. Unless one understands concepts like means and standard deviations, d would require further explanation, and it would be even more difficult to explain the correlation between sex and height. Further, these figures assume that equal-size groups are obtained in one's sample. Setting aside the influence of normal sampling error, unequal group size alone could cause d to range from 1.93 (if nearly all participants are men) to 2.08 (if nearly all participants are women); r_{pb} could range from a high of .71 (with approximately equal numbers of men and women) to arbitrarily close to .00 (if nearly all participants are of the same sex). In contrast, $A = .92$ regardless of group sizes, and it is fairly easy to understand what this means: Selecting pairs of men and women at random, there is a 92% chance that the man is taller.

APPLICATIONS IN OTHER RESEARCH CONTEXTS

Based on the many potential advantages of a probability-based effect size statistic, McGraw and Wong (1992) proposed ways to use the *CL* statistic for use with discrete-valued variables and in studies with more than two groups or correlated samples. Vargha and Delaney (2000) expressed their enthusiasm for McGraw and Wong's innovative work but also described some weaknesses of the *CL* statistic and its generalizations. As noted earlier, the calculation of *CL* and its offshoots requires parametric assumptions. The nonparametric A statistic is not constrained in this way. More important, the proposed extensions differ conceptually from the original *CL* statistic and are not generalizations in that they do not contain it as a special case when there are $k = 2$ groups. A related limitation is that the extensions of the *CL* statistic do not share its interpretation: a value of .50 does not always correspond to stochastic equality. Expanding in similar directions, Vargha and Delaney (2000) introduced their own generalizations of the *A* statistic that do not share these weaknesses. We review their generalizations and introduce some of our own to afford researchers even greater flexibility in adapting this versatile statistic to meet their needs. We begin with a straightforward way to handle discrete data and then discuss generalizations of *A* for use with weighted data, $k > 2$ groups, and correlated samples. Each of these applications contains an empirical illustration using published data. Once the full range of generalizations and extensions has been presented, we describe and illustrate a method that can be used to construct confidence intervals for each application of *A*. All calculations were performed using a suite of programs written in R that we make available at http://www.tcnj.edu/∼ruscio/taxometrics.html

The Discrete Values Case

McGraw and Wong's (1992) *CL* statistic was designed for use with continuous, rather than discrete, data. Although they proposed a way to adapt *CL* for use with discrete values (e.g., dichotomous or multinomial data), Vargha and Delaney (2000) discussed a number of conceptual weaknesses with their approach and, by way of a better alternative, noted that the *A* statistic requires no modification for use with discrete values because it allows for tied scores. Moreover, the *A* statistic can be useful in contexts in which the parametric assumptions underlying other effect size measures, such as *d* or r_{pb}, would be violated by the discreteness of the data (e.g., data taking a small number of discrete values cannot approximate a normal distribution especially well). All of the data that we analyze throughout this article vary along discrete scales with varying numbers of unique scores.

Empirical illustration. We reanalyzed data from Conley (2011) to illustrate the use of *A* as an effect size statistic in the discrete values case. Conley studied gender differences in the acceptance of casual sex offers. In a hypothetical scenario, 516 individuals were offered sex by a stranger and asked on a 7-point scale (1 = *not at all likely* to 7 = *extremely likely*) how likely they would be to accept the offer. Conley classified those who responded with a 1 as definitely rejecting the offer (which included 82% of women) and those who responded with 2 through 7 as entertaining the possibility of the offer (which included 74% of men); no explicit rationale for the decision to cut the scores at 2 was provided. In the table of gender comparisons, scores on the full 7-point scale were retained for analysis, and there was a substantial gender difference ($M = 3.74$, $SD = 2.16$

for men; $M = 1.37$, $SD = 0.97$ for women). Because scores on a 7-point scale are not continuous, the distributions were nonnormal, and the variances were unequal to a potentially problematic extent, it is questionable whether Cohen's *d* is an appropriate effect size estimator in this instance. Calculating *A* instead yields a value of .82, indicating an 82% chance that a randomly chosen man would respond more favorably to an offer of casual sex than a randomly chosen woman. This nonparametric statistic is appropriate for these data and yields a simple, easy-to-understand estimate of the magnitude of the gender difference.

The Weighted Data Case

Social and behavioral scientists occasionally analyze data with differential weights assigned to cases. For example, epidemiologists might weight data across racial or ethnic groups to obtain statistics that can be generalized to a national population even when some subgroups are over- or underrepresented (e.g., by design or due to normal sampling error). The *A* statistic can be calculated for weighted data in a straightforward manner: each time a score comparison is made, weight the credit in the numerator (e.g., 1 if the score for the member of Y_1 is larger than the score for the member of Y_2, .5 for a tie) by the product of these two cases' weights. Likewise, increment the denominator by the product of the weights rather than by 1 unit. This gives each comparison the weight jointly merited by the two cases' individual weights, and the resulting weighted version of the *A* statistic will therefore take all weights into account. Expressed as an equation, the weighted version of *A* is calculated as follows:

$$A = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{1i}w_{2j}([Y_{1i} > Y_{2j}] + .5[Y_{1i} = Y_{2j}])}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{1i}w_{2j}}, \qquad (2)$$

where w_{1i} is the weight for the *i*th member of group Y_1, w_{2j} is the weight for the *j*th member of group Y_2, Y_{1i} refers to the score for the *i*th member of group Y_1, Y_{2j} refers to the score for the *j*th member of group Y_2, the expression $[Y_{1i} > Y_{2j}]$ is scored as 1 if true and 0 otherwise, and the expression $[Y_{1i} = Y_{2j}]$ is scored as 1 if true and 0 otherwise. The *A* statistic as originally described can be considered a special case in which all cases of data are assigned equal weights, in which case Equation 2 simplifies to Equation 1.

Applying weights in this way does not affect the range of values that *A* can take (.00 to 1.00). Because the denominator is incremented using the same product of case weights as the numerator for each pairwise comparison, multiplying or dividing all case weights by a constant will not affect the value of *A*. This means that it is unnecessary for the weights to sum to 1 (or any other specific value). This method for applying case weights can be generalized to all variants of the A statistic described in this article, and each of our computer programs allows users to provide case weights.

Empirical illustration. We reanalyzed data from Buturovic and Klein (2010) to illustrate the use of the A statistic with weighted data. Buturovic and Klein studied economic enlightenment in a nationwide telephone survey of 4,835 American adults. Their primary dependent variable was the number of items (out of eight) for which each participant's selected responses were considered "unenlightened" when evaluated against the consensus of opinion among economists based on theoretical and empirical scholarship (e.g., disagreeing with the statement "Rent control leads to housing shortages" was scored as unenlightened; see Jenkins, 2009). We reverse-scored this measure so that higher values represented greater economic enlightenment. There was a gender difference such that men scored a bit higher ($M = 5.40$, $SD = 2.22$) than women ($M = 4.42$, $SD = 2.18$), but this may not generalize well to the U.S. population because (a) there was also a large difference in economic enlightenment by self-identified political party affiliations (Ms = 3.41, 6.39, and 4.97 for Democrats, Republicans, and Independents, respectively), (b) the sample was not representative of the U.S. population with respect to party affiliation (according to Gallup surveys in 2010), and (c) the sample contained different proportions of women in each party (50%, 35%, 33% for Democrats, Republicans, and Independents, respectively). To address these concerns, weights were applied to adjust for discrepancies between party affiliations in the sample and population. For example, when a score comparison was made between a Democrat and a Republican, the result was multiplied by $0.853 \times 0.712 = 0.607$, a low value because both parties were overrepresented in the sample; when a score comparison was made between two Independents, the result was multiplied by $1.812 \times 1.812 = 3.283$, a much larger value because Independents were underrepresented in the sample. Among the 4,519 individuals who reported their gender and were affiliated with one of these three political parties, adjusting all pairwise comparisons between men and women using these weights yielded $A = .62$. This estimate suggests that in the U.S. population, a randomly selected man would have a 62% chance of scoring higher on Buturovic and Klein's measure of economic enlightenment than a randomly selected woman.

The *k*-Groups Case

Vargha and Delaney (2000) introduced two generalizations of the A statistic for use with more than two groups, both of which are analogous to an omnibus F statistic in that they quantify the extent of differences observed among all groups.

Vargha and Delaney referred to their first generalization of A as an estimator of *stochastic homogeneity*. In this first generalization, the A statistic is used to determine whether scores in one group differ from those in the union of all other groups, and this procedure is then repeated for each group in turn. The resulting series of A values are then aggregated into a single statistic by calculating the average absolute deviation (AAD) from .50 (the value representing stochastic equality). Because stochastic homogeneity would yield AAD = .00, Vargha and Delaney noted that one can add .50 to return this measure to the original scale of the A statistic. We label this variant of A as A_{AAD} because it is based on the AAD. The population value Δ_{AAD} is estimated as follows:

$$A_{AAD} = \frac{\sum_{i=1}^{k} |A_{ik} - .50|}{k} + .50,$$

where k is the number of groups and A_{ik} is the value of the A statistic when comparing group i with the union of all other groups.

Vargha and Delaney (2000) referred to their second generalization of A as a measure of *pairwise stochastic equality*. In this second generalization, the A statistic is calculated for all pairs of groups, and these values are aggregated by calculating the average absolute pairwise deviation (AAPD) from .50. Once again, .50 can be added to the resulting value to return it to the scale of A. We label this variant of A as A_{AAPD} because it is based on the AAPD. The population value Δ_{AAPD} is estimated as follows:

$$A_{AAPD} = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |A_{ij} - .50|}{\frac{k(k-1)}{2}} + .50,$$

where A_{ij} is the value of the A statistic when comparing groups i and j.

Interested readers can consult Vargha and Delaney (2000) for further details regarding the calculation of A_{AAD} and A_{AAPD}. As noted earlier, both of these statistics are nonparametric and contain the original formulation of A as a special case when $k = 2$ groups are compared. Although they yield values on the same scale as A (from .50, stochastic equality, to 1.00, no overlap between any scores), their interpretation is a little more complex. A relatively minor complication is that neither of these generalizations can fall below .50 because of the use of absolute deviations to aggregate values. Because neither AAD nor AAPD can be negative, adding .50 to rescale them into A_{AAD} and A_{AAPD} ensures a minimum value of .50 for each. The greater difficulty in interpreting these estimators is analogous to what one encounters with any omnibus statistic: in this case, specifically, one cannot determine which groups' distributions differ from which other groups' distributions. One potentially helpful adjunct to these generalizations would be to use the original A statistic to compare specific pairs of groups in a manner analogous to performing post hoc comparisons of means following an F test.

In addition to generalizations of A that correspond to omnibus statistics or specific pairwise comparisons, one can construct variants that impose different constraints on the group comparisons. First, one can single out a particular group for comparison with all others, pooling the latter as though drawn from a single population. The method of calculation is straightforward: treat scores from all but the reference group as though drawn from a single population, and compare them with the scores from the reference group using the A statistic as usual. We label this variant of A as A_{ik} to signify the comparison of one group with all of the others. The A_{ik} statistic estimates a population value that can be expressed in this way: $\Delta_{ik} = \Pr(Y_i > Y_{\sim i})$, where $\sim i$ denotes the union of all groups except i. The probability is estimated by calculating A as shown in Equation 1. Second, one can quantify the extent to which scores tend to be rank-ordered across groups. To calculate this variant, one first calculates A for each pair of adjacent groups and then takes the mean of these values. We label this as A_{ord} to signify the generalization of A to an ordinal comparison across multiple groups, and A_{ord} estimates the population value defined as follows: $\Delta_{ord} = [\Pr(Y_1 > Y_2) + \Pr(Y_2 > Y_3) + \ldots + \Pr(Y_{(k-1)} > Y_k)]/[k-1]$. Each probability in this expression is estimated using the formula for A shown in Equation 1.

Empirical illustration. We reanalyzed data from Orth, Robins, and Soto (2010) to illustrate each generalization of A as an effect size statistic for comparisons among more than two groups. Using cross-sectional data from 2,611 individuals, Orth et al. tracked changes in emotions across the life span. They found that Authentic Pride exhibited a monotonically increasing trend over the seven age groups in the study. Because there were so many groups, calculating an effect size estimator such as d for all pairwise comparisons, or even for all nearest-neighbor pairs of age groups, would have been unwieldy. Fortunately, generalizations of A to the k-groups case can be helpful in this context, and all four options are illustrated here.

Vargha and Delaney's (2000) versions yielded $A_{AAD} = .56$ and $A_{AAPD} = .58$, both of which suggest fairly small—but not necessarily negligible—differences among age groups. It remains unclear, however, which groups differ more or less from others. Singling out the youngest age group for comparison against all others yielded $A_{ik} = .45$ (meaning that this age group exhibited a bit less Authentic Pride than did the composite of all other age groups), whereas singling out the oldest age group for comparison yielded $A_{ik} = .63$ (more Authentic

Pride than in other age groups), a larger effect. The monotonic trend across all age groups was quantified by a value of $A_{ord} = .53$. Although this result suggests a rather slight difference, it must be remembered that A_{ord} compares change across adjacent age groups only. The probability that members of the oldest age group exhibited greater levels of Authentic Pride than members of the youngest age group was $A = .66$, which suggests that a considerably larger effect accumulates over a longer developmental span.

The Correlated Samples Case

The A statistic can be generalized to applications with correlated samples (e.g., repeated measures, matched samples). The key to these variants is to compare scores across measures for each case rather than making all pairwise score comparisons across groups. Vargha and Delaney (2000) introduced this approach for the case of $k = 2$ correlated samples as follows:

$$A = [\#(Y_1 > Y_2) + .5\#(Y_1 = Y_2)]/n,$$

where Y_1 and Y_2 now refer to scores on two measures and comparisons are made across n participants. When calculated in this way, A represents the chance that a randomly chosen participant's score would be greater on the first than the second measure, allowing partial credit for ties. Formally, A still serves an estimator of $\Delta = \Pr(Y_1 > Y_2)$, with Y_1 and Y_2 representing correlated samples rather than discrete groups.

Not only can A be extended for use with $k = 2$ correlated samples but also this application can be generalized to $k > 2$ correlated samples in the same four ways that were described for $k > 2$ groups. For Vargha and Delaney's (2000) variants A_{AAD} and A_{AAPD} as well as the A_{ik} and A_{ord} variants introduced here, the key remains making score comparisons across measures within cases rather than pairwise across groups. For each of our programs that calculates a k-groups variant of A, there is a parallel program that calculates a variant of A for k correlated samples.

Empirical illustration. We reanalyzed data from Peetz and Kammrath (2011) to illustrate the use of the A statistic with correlated samples. Peetz and Kammrath studied the number of promises that were made ($M = 2.77$, $SD = 0.53$) and that were subsequently broken ($M = 2.23$, $SD = 1.02$) by 83 individuals in interpersonal relationships, noting that there was a statistically significant difference ($t[82] = 5.18$, $p = .001$) but not including an effect size estimate. Calculating the correlated-samples variant of A yielded $A = .66$, indicating a 66% chance that a randomly chosen individual kept fewer promises than he or she had made. Though Peetz and Kammrath did not choose to report results in this way, one could also compare the number of promises kept with the number broken. This yielded $A = .83$, indicating an 83% chance that a randomly chosen individual kept more promises than he or she broke.

Constructing Confidence Intervals

In the APA's (2009) *Publication Manual*, researchers are asked to report appropriate effect size statistics for their analyses. To enhance the utility of the effect size estimate, the APA recommends accompanying it with a confidence interval (CI) to indicate the precision of the estimated effect size (APA, 2009). A number of methods have been proposed for constructing CIs around the *A* statistic. Ruscio and Mullen (2012) examined the performance of nine analytic methods (e.g., estimating a standard error [*SE*] of a theoretical sampling distribution and constructing a 95% CI as $A \pm 1.96 \times SE$) and three bootstrap methods (e.g., resampling with replacement to generate an empirical sampling distribution of the *A* statistic and constructing a 95% CI such that it spans the middle 95% of the observed values). Based on a variety of criteria, they recommended using the bias-corrected and accelerated (BCA) bootstrap method (Efron & Tibshirani, 1993) to construct CIs for *A*. Our suite of programs allows users to implement the BCA bootstrap method to construct a CI for any of the generalizations of *A* described in this article. In addition, these programs provide an estimate of the *SE* calculated as the *SD* of all bootstrap samples' values of *A*.

Empirical illustration. Each of the analyses performed in this article was updated using $B = 1{,}999$ bootstrap samples to construct a 95% CI using the BCA method. For the discrete data case using Conley's (2011) data, there was an 82% chance that a randomly chosen man would be more accepting of an offer of casual sex than a randomly chosen woman, with the CI for $A = (.779, .853)$. The asymmetry of this CI—it extends a little further from .82 on the lower than the upper end—is not uncommon when *A* departs from .50 and is part of the reason that the BCA method of CI construction outperformed many of the alternatives. Unlike most analytic methods for constructing CIs, certain bootstrap methods (including the BCA method) can yield asymmetric CIs when the sampling distribution of a statistic is itself asymmetric. The fact that the CI does not include .50, which represents stochastic equality, suggests that one would reject this null hypothesis (for more on hypothesis testing using *A* and its generalizations, see Vargha & Delaney, 2000).

For the weighted data case using Buturovic and Klein's (2010) data, with sample weights applied to represent party affiliations in the U.S. population, there was a 62% chance that a randomly selected man would score higher on economic enlightenment than a randomly selected woman, CI = (.596, .629). Because of the very large sample size ($N = 4{,}519$), this is an especially narrow CI.

For the *k*-groups case using Orth et al.'s (2010) data, differences in Authentic Pride across age groups were quantified in many ways. When all groups were compared, the differences were modest, with $A_{AAD} = .56$, CI = (.545, .570) and $A_{AAPD} = .58$, CI = (.559, .590). Here, too, a large sample size ($N = 2{,}601$) provided excellent precision in estimating the size of the effect. Singling out the youngest age group for comparison against all others yielded $A_{ik} = .45$, CI = (.428, .485), whereas singling out the oldest age group for comparison yielded $A_{ik} = .63$, CI = (.590, .673). Quantifying the monotonic trend across all age groups yielded $A_{ord} = .53$, CI = (.518, .535), and contrasting members of the youngest and oldest age groups yielded $A = .66$, CI = (.615, .708). None of these CIs includes .50, suggesting that even these values of *A* that represent very small effects would be statistically significant with two-tailed tests at $\alpha = .05$.

Finally, for the correlated samples case using Peetz and Kammrath's (2011) data, there was a 66% chance that a randomly chosen individual kept fewer promises than he or she had made, CI = (.596, .711), and an 83% chance that a randomly chosen individual kept more promises than he or she broke, CI = (.729, .892). The latter CI, which deviates farther from .50 than any other CI reported here, demonstrates the greatest degree of asymmetry as well.

CONCLUSIONS

We reviewed generalizations of a probability-based estimator of effect size beyond its original application in two-group comparisons. Specifically, we showed how to use the *A* statistic with discrete data, weighted data, $k > 2$ groups, and correlated samples. The use of each generalization was illustrated through reanalyses of data from published studies. Following the advice in the APA's (2009) *Publication Manual* to accompany effect size estimates with CIs, we drew from related research to include a bootstrap method for constructing CIs for each generalization of the *A* statistic discussed in this article. All of our calculations were performed using a suite of free programs that should make it easy to use the *A* statistic and accompany it with a CI in a wide variety of research contexts.

ACKNOWLEDGMENTS

We are grateful to Terri Conley, Ulrich Orth, and Johanna Peetz for providing us with their data sets and permission to reanalyze them for illustrative purposes. Buturovic and Klein (2010) made their data available through a link in their original article, and we appreciate this as well.

REFERENCES

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

Buturovic, Z., & Klein, D. B. (2010). Economic enlightenment in relation to college-going, ideology, and other variables: A Zogby survey of Americans. *Econ Journal Watch, 7,* 174–196.

Conley, T. D. (2011). Perceived proposer personality characteristics and gender differences in acceptance of casual sex offers. *Journal of Personality and Social Psychology, 100,* 309–329. doi: 10.1037/a0022152

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.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79,* 314–316.

Grisson, 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. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge.

Hsu, L. M. (2004). Biases of success rate differences shown in Binomial Effect Size Displays. *Psychological Methods, 9,* 183–197.

Jenkins, B. (2009). Rent control: Do economists agree? *Econ Journal Watch, 6,* 73–112.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759. doi: 10.1177/0013164496056005002

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361–365. doi: 10.1037/0033-2909.111.2.361

Orth, U., Robins, R. R., & Soto, C. J. (2010). Tracking the trajectory of shame, guilt, and pride across the life span. *Journal of Personality and Social Psychology*, *99,* 1061–1071. doi: 10.1037/a0021342

Peetz, J., & Kammrath, L. (2011). Only because I love you: Why people make and why they break promises in romantic relationships. *Journal of Personality and Social Psychology*. Advance online publication. doi: 10.1037/a0021857

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13,* 19–30. doi: 10.1037/1082-989X.13.1.19

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research, 47,* 201–223. doi: 10.1080/00273171.2012.658329

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. doi: 10.3102/10769986025002101