A robust effect size measure $A_w$ for MANOVA with non-normal and non-homogenous data

Johnson Ching-Hong Li¹, Marcello Nesca², Rory Michael Waisman³, Yongtian Cheng¹ and Virginia Man Chung Tze⁴

Abstract
A common research question in psychology entails examining whether significant group differences (e.g., male and female) can be found in a list of numeric variables that measure the same underlying construct (e.g., intelligence). Researchers often use a multivariate analysis of variance (MANOVA), which is based on conventional null-hypothesis significance testing (NHST). Recently, a number of quantitative researchers have suggested reporting an effect size measure (ES) in this research scenario because of the perceived shortcomings of NHST. Thus, a number of MANOVA ESs have been proposed (e.g., generalized eta squared $\eta^2$, generalized omega squared $\omega^2$), but they rely on two key assumptions—multivariate normality and homogeneity of covariance matrices—which are frequently violated in psychological research. To solve this problem we propose a non-parametric (or assumptions-free) ES ($A_w$) for MANOVA. The new ES is developed on the basis of the non-parametric $A$ in ANOVA. To test $A_w$ we conducted a Monte-Carlo simulation. The results showed that $A_w$ was accurate (robust) across different manipulated conditions—including non-normal distributions, unequal covariance matrices between groups, total sample sizes, sample size ratios, true ES values, and numbers of dependent variables—thereby providing empirical evidence supporting the use of $A_w$, particularly when key assumptions are violated. Implications of the proposed $A_w$ for psychological research and other disciplines are also discussed.

Keywords
Effect size, MANOVA, robust statistics, Monte-Carlo simulation, experimental design

Effect size (ES), a quantity that directly presents or measures the strength of an effect in a study, has received increasing attention. ES is regarded as a supplement to conventional null hypothesis significance testing (NHST) because NHST focuses on making a dichotomized decision to reject or accept a research hypothesis without considering the level or magnitude of the effect observed in a study. In fact, many methodologists, professional associations, and journal editors have suggested that researchers should report ESs in addition to NHST (Murphy, 1997; Thompson, 1994; Trafimow and Marks, 2015). The American Psychological Association (APA) publication manual is more assertive on this matter, stating: “[a]void relying solely on statistical hypothesis testing, such as $P$ values, which fails to convey important information about effect size.” (Mathews and Mathews, 2007, Section IV.A.6.c).
In light of the call for reporting ESs, different estimates of the true ES in the multivariate analysis of variance (MANOVA) have been developed in the literature. For example, Steyn and Ellis (2009) evaluated the accuracy of conventional MANOVA ESs (e.g. generalized eta squared $\eta_\Lambda^2$, generalized omega squared $\omega_\Lambda^2$, etc.) when the key assumptions—multivariate normality and homogeneity of covariance matrices—are met. Other researchers (e.g. Finch, 2016; Fouladi and Yockey, 2002; Olejnik and Huberty, 2006) also discussed the conceptual and computational details for these MANOVA ESs. However, both of these ESs do not appear to be robust to violations of these assumptions because their calculations are developed based on their corresponding, univariate estimates (i.e. eta squared ($\eta^2$) and omega squared ($\omega^2$)) that were found to be biased ES estimators when these assumptions were violated (e.g. Troncoso Skidmore and Thompson, 2013). Unfortunately, these assumptions are frequently violated in behavioral and social sciences research (e.g. Algina et al., 2005). Hence, there is an urgent need for developing a more appropriate MANOVA ES that does not rely on these assumptions.

This study aims to develop a more robust ES measure ($A_w$) and compare it with the conventional ESs (i.e. $\eta_\Lambda^2$ and $\omega_\Lambda^2$) for MANOVA when the parametric assumptions (i.e. multivariate normality and homogeneity of covariance matrices) are violated. The proposed $A_w$ is based on Vargha and Delaney’s (2000) development of an ES measure for univariate analysis of variance (i.e. $A$) with only one dependent variable (DV), which is a more restricted case of MANOVA.

This paper is divided into six sections. The first section explains the importance of the two assumptions in MANOVA ESs. The second section presents the computational details of the conventional (parametric) ESs (i.e. $\eta_\Lambda^2$ and $\omega_\Lambda^2$) that depend upon the two assumptions. The third section discusses the development of the non-parametric $A_w$ and its parametric counterpart $CL_w$ (called common language ES). In the fourth section, methods and design of the Monte-Carlo simulation study are described. The fifth section presents and discusses the simulation findings. In the sixth section, conclusion and implications of the findings are discussed.

**Two key assumptions: Multivariate normality and homogeneity of covariance matrices**

In most commercial statistical packages (e.g. SPSS), four conventional NHST statistics for MANOVA are typically reported: Pillai’s Trace, Wilks’ Lambda, Hotelling-Lawley Trace, and Roy’s Greatest Root. These statistics are known to be reliable when two key assumptions are met: multivariate normality and homogeneity of covariance matrices. Multivariate normality means that a vector of weighted DV scores is independently and normally distributed for each level of the IV. Homogeneity of covariance matrices requires that the variance and covariance of the DVs are the same for each level of the IV. However, data in behavioral science often deviates from these assumptions (Keselman and Lix, 1997). Researchers in behavioral science often use single- or multi-item measures that employ Likert-scale (e.g. total score of three items on a 5-point scale), and hence, the boundaries of the total scores are fixed (e.g. 3–15 points) producing a non-normal (i.e. platykurtic) distribution. Data observed in groups (e.g. clinical patients, gifted children) also tend to follow a heavy-tailed (i.e. skewed) distribution. Furthermore, the assumption of homogeneity of covariance matrices is rarely met in behavioral research (Tang and Algina, 1993). For example, data in a clinical group tend to have a smaller variance than in a normal group. A treatment may also be effective in systematically improving the outcome of interest among treated participants, and so they appear to be more homogenous (i.e. ceiling effect) than participants receiving no treatment.

Despite the popularity of the conventional statistics in reporting MANOVA results, previous research has found that they are not robust to violations of these important assumptions. Everitt (1979) found that these statistics tend to have an inflated Type II error (i.e. fail to reject the null hypothesis when it is false) when the degree of skewness in the DVs increases. Algina et al. (1991) found that when data are asymmetrically non-normal, the test statistics lead to inflated Type I error, especially when the covariances are heterogeneous and the sample size is unbalanced between groups. Cole et al. (1994) showed that the performance of these statistics varied. When the off-diagonal elements in the covariance matrices increase in difference across groups, the test statistics become less robust. Hopkins and Clay (1963) showed that when the covariance matrices differ between two groups, the test statistics are unlikely to be robust. Given that the conventional MANOVA ESs are based on Wilk’s Lambda, one of the conventional test statistics for MANOVA, these ESs are expected to be affected by the same assumption violations. However, no study has systematically evaluated the performance of these ES, nor has prior work proposed a non-parametric ES that does not rely on these assumptions. Here we report the first study to do both.

**Conventional ESs**

In two-independent samples univariate analysis of variance (ANOVA), there are two common families of ESs: the difference ($d$) family and correlation ($r$) family. For the $d$-family, the standardized mean difference between two groups in DV scores can be used to evaluate the strength of the effect. That is, $d = (\bar{X}_p - \bar{X}_q) / s_{pooled}$, where the numerator refers to the mean difference between Groups $p$ and $q$, and $s_{pooled}$ is the pooled standard deviation. For the $r$-family, ES can be expressed in terms of Pearson’s correlation ratio, which is defined as the proportion of total variability that can be explained by an IV. That is, $\eta^2 = SS_r / SS_{total}$, where $SS_r$ is the sum of squares due to the IV, and $SS_{total}$ is the sum of squares of the DV scores without the IV. $\eta^2$ is a measure of model goodness-of-fit and is conceptually equivalent to $R^2$ in
multiple regression, and hence, it is regarded as a measure of association between the IV and DV. Note that $\eta^2$ is mathematically related to $d$, that is, 
$$
\eta^2 = \frac{d^2}{d^2 + \frac{(n_1 + n_2)^2}{n_1 n_2}}.
$$

**Generalized eta squared $\eta^2_\Lambda$**

In MANOVA, the generalized eta squared $\eta^2_\Lambda$ is defined as the total generalized variance that can be explained by the between-group generalized variance, that is,

$$
\eta^2_\Lambda = 1 - \Lambda = 1 - \left| \frac{\sum_{j=1}^{J} |\sum_j w_j (\mu_j - \mu)| \sum_j w_j + \sum_{\theta} \right| \tag{1}
$$

where $\Lambda$ is the Wilk’s lambda, which consists of two components. First,

$$
\sum_B = \left( \frac{1}{N} \right) \sum_{j=1}^{J} N_j (\mu_j - \mu) (\mu_j - \mu),
$$

is the between-groups covariance matrix, $N$ is the total sample size, $N_j$ is the sample size for $j=1, 2, \ldots, J$ groups, $\mu_j$ is the mean for the $j$th group, and $\mu$ is the grand mean. Second,

$$
\sum_W = \left( \frac{1}{N} \right) \sum_{j=1}^{J} \sum_{n=1}^{N} (y_{jn} - \mu_j) (y_{jn} - \mu_j),
$$

is the within-groups covariance matrix, where $y_{jn}$ is the $n$th observation for the $j$th group. According to Cohen (1988), $\eta^2_\Lambda$ is similar to $R^2$ in multiple regression, which measures the association between the weighted level or set of the DVs and the IV in MANOVA.

**Generalized omega squared $\omega^2_\Lambda$**

The eta squared ($\eta^2$) in ANOVA is commonly regarded as a biased estimator for the true ES—it tends to overestimate the true ES (Mordkoff, 2019). Researchers have proposed and developed another ES, called omega squared ($\omega^2$), for ANOVA, which can adjust for the bias found in $\eta^2$. The multivariate generalization of $\omega^2$ is known as the generalized omega squared,

$$
\omega^2_\Lambda = 1 - \left( \frac{NA}{N-J+\Lambda} \right) \tag{2}
$$

where $J$ is the number of groups in the IV, and $N$ and $\Lambda$ are defined in equation (1).\(^1\)

**Recently developed ESs**

**Parametric common language ES (CL) in ANOVA**

The idea of common language ES (CL) can be found in statistical studies published almost 80 years ago when Wilcoxon (1945) proposed a rank-order comparison for scores observed between two treatments. Mann and Whitney (1947) extended Wilcoxon’s method, which calculated the rank numbers in which the scores in treatment $A$ are larger than the scores in treatment $B$, by further defining the statistical properties (e.g. probability distribution) of Wilcoxon’s statistical measure, known as the $U$ statistic. Govindarajulu (1967) was one of the early studies that formally defined the meaning of a probability estimate, $P(X < Y)$, where $X$ refer to continuous scores in a random sample $A$ (e.g. treatment), and $Y$ refer to continuous scores in a random sample $B$ (e.g. control). This measure quantifies the probability that a score in one sample is stochastically smaller than a score in another sample. Govindarazulu focused on deriving an analytic method that constructs the confidence intervals surrounding the measure of $P(X < Y)$. Wolfe and Hogg (1971) published a tutorial paper that encourages the use of $P(X < Y)$ among applied statisticians and researchers.

Based on these early papers published in statistics journals, McGraw and Wong (1992) was one of the pioneer studies in psychology that proposed the use of a statistic that measures “how often a score sampled from one distribution will be greater than a score sampled from another distribution” (p. 361). They labeled this statistic a common language ES (CL) and proposed its use as a type of probability of superiority statistic for univariate ANOVA. CL estimates the parameter, $P(Y_p > Y_q)$, which measures the probability that a score in group $p$ is higher than a score in group $q$. For example, if a researcher is comparing the effect of an intervention group relative to a control group, the researcher can present the CL that estimates the probability (e.g. 80% chance) that the observations (e.g. self-esteem) would be better for a randomly selected member of the intervention group than for a randomly selected member of the control group. Hsu (2004) regarded this as a more intuitive way to conceptualize ES, as it is easy for researchers and practitioners to understand even without formal training in statistics. According to Ruscio (2008), when data meet the parametric assumptions (i.e. normality and homogeneity of variances), CL can be expressed as

$$
CL = \Phi \left[ \frac{\bar{X}_p - \bar{X}_q}{s_{pooled}} \right],
$$

where $\Phi$ is the normal cumulative distribution function, $\bar{X}_p$ and $\bar{X}_q$ are the means of Groups $p$ and $q$, respectively, and $s_{pooled} = \sqrt{\left( \frac{(N_p-1)s_p^2 + (N_q-1)s_q^2}{N_p + N_q - 2} \right)}$ is the pooled SD of the DV scores for the two groups.

**Non-parametric A in ANOVA**

Vargha and Delaney (2000) criticized McGraw and Wong’s (1992) CL on the basis that it assumes the scores follow a normal distribution. To overcome this problem they proposed a non-parametric estimator, known as the measure of stochastic superiority (i.e. $A$, equation (2) in Vargha and Delaney, 2000), for use in ANOVA with two independent samples. $A$ measures effect size based on the probability that a score in group $p$ is higher than a score in group $q$, that is,
\[ A = \frac{\#(p > q) + 0.5\#(p = q)}{n_p n_q}, \]  

(3)

where \# is the count function, \( p \) and \( q \) are vectors of scores for the two samples, and \( n_j \) is the sample size in group \( j=p, q \). Assume \( p = \{5, 7, 6, 5\} \) and \( q = \{3, 5, 5, 1\} \), the count function \((i.e. \#(p = 5) = \{q = 3, 5, 5, 1\})\) yields a total count of 3. Repeat this process for the remaining elements in \( p, A = (3 + 4 + 4 + 3)/16 = 0.875 \), meaning that there is a 87.5% chance that the observation would be higher for a randomly selected member of group \( p \) than for a randomly selected member of group \( q \).

**Parametric CL\(_w\) and non-parametric \( A_w\) in MANOVA**

To generalize \( A \) in two independent-groups ANOVA to the multivariate complement (i.e. \( A_w \)) in MANOVA, we can substitute the \( w_p \) and \( w_q \) vectors of scores—a linear composite of two or more DVs for each participant in group \( p \) and \( q \), respectively—into the \( p \) and \( q \) vectors of scores in equation (3). Assume \( X^{(j)} \) is the \( N \) (number of participants) by \( p \) (number of DVs) data matrix for group \( j=1 \) and 2. \( X^{(j)} = X_1^{(j)} , X_2^{(j)} , ..., X_v^{(j)} \), where \( X_v^{(j)} \) is the vector of scores of the \( v \)th DV in group \( j \). The linear composite becomes, \( w_p = a_1^{(j)} X_1^{(j)} + a_2^{(j)} X_2^{(j)} + ... + a_{v}^{(j)} X_v^{(j)} \), where, \( a_1^{(j)}, a_2^{(j)}, ..., a_{v}^{(j)} \) are the weights (or eigenvectors) for score vectors \( X_1^{(j)}, X_2^{(j)}, ..., X_v^{(j)} \) for group \( j=1 \) and 2 based on discriminant function analysis. The vector of \( a \) weights can be estimated by

\[ a = s_{pooled}^{-1}(X_p - X_q), \]

(4)

where \( X_p \) and \( X_q \) are the \( v \) by 1 vectors that contains the means of the DVs for group \( j=p \) and \( q \), respectively, and \( s_{pooled}^{-1} = (SSCP_p + SSCP_q)/(N_p + N_q - 2) \) with \( SSCP_p \) and \( SSCP_q \) are the sum of squares and cross products matrices for group \( j=p \) and \( q \), respectively. These weights are selected to produce the maximum possible \( (w_p - w_q) \) difference, which can be obtained in a statistical package (e.g. SPSS). Hence, \( A_w \) can be expressed as

\[ A_w = \frac{\#(w_p > w_q) + 0.5\#(w_p = w_q)}{n_p n_q}, \]

(5)

which expresses the probability (e.g. 90%) that the linear composite \( w_p \) would be higher for a randomly selected member of group \( p \) than for a randomly selection member of group \( q \). The parametric estimator for the probability of superiority in MANOVA (\( CL_w \)) can be expressed as

\[ CL_w = \Phi \left[ \frac{(X_p - X_q)}{s_w} \right], \]

(6)

where \( X_p \) and \( X_q \) are the means of the weighted composite scores for groups \( p \) and \( q \), respectively, and \( s_w \) is the pooled SD of the weighted composite scores from the two groups.

This study evaluates the performance of two conventional parametric ESs—\( \eta^2_w \) (equation (1)) and \( \omega^2_w \) (equation (2))—and two probability-based ESs—non-parametric \( A_w \) (equation (5)) and parametric \( CL_w \) (equation (6)).

**Monte-Carlo study**

**Design**

Seven factors that would affect the performance of \( \eta^2_w \), \( \omega^2_w \), \( CL_w \), and \( A_w \) were evaluated.

**Factor 1: Standardized mean vector difference (\( \delta \); four levels).** This parameter reflects the level of standardized mean difference between the weighted DV scores in two groups, which is similar to Cohen’s standardized mean difference \( d \) in ANOVA. According to Cohen (1988), in social science research a \( d \) of 0.20, 0.50, and 0.80 reflects a small, moderate and large ES respectively. In addition, the value of 1.50 was included to examine the impact of an extremely strong \( d \) on the observed ESs. The corresponding values for \( \eta^2_w \) are 0.0099, 0.0588, 0.1379, and 0.3600, and for \( CL_w \) and \( A_w \) are 0.5793, 0.6915, 0.7881, and 0.9332. The corresponding true values for \( \omega^2_w \) are similar but slightly deviated from the true values for \( \eta^2_w \). The adjustment depends on two factors: the sample size (\( N \)) and number of groups (\( J \)). That is, combining equations (1) and (2), \( \omega^2_w \) is adjusted by:

\[ \omega^2_w = 1 - \left[ \frac{N(1-\eta^2_w)}{N - J + (1-\eta^2_w)} \right]. \]

Hence, taking a simulated condition with the total sample size of 150 (i.e. \( N=150 \)) and the number of groups equals 2 (i.e. \( J=2 \)) as an example, when the value for \( \eta^2_w \) was 0.0099, the adjusted value for \( \omega^2_w \) was estimated to be 0.0032 (i.e. \( 1 - \left[ 150(1-0.0099) \right] / 150 - 2 + (1-0.0099) = 0.0032 \)).

**Factor 2: Distribution (\( \Theta \); six levels).** In addition to the normal distribution (\( N(0,1) \) with skewness (\( \gamma_1 \))=0 and kurtosis (\( \gamma_2 \))=0), five non-normal (i.e. two peaked, two skewed, and one mixed normal) distributions were generated based on Algina et al. (2005). The peaked distribution is characterized by a long (or short) tail with few (or most) scores clustered around the center of the distribution. The skewed distribution is characterized by unequal-length tails along a distribution. The mixed normal distribution appears to be a normal distribution but with longer tails on both ends, mimicking a distribution with outliers on both ends. Following Algina et al., for the peaked and skewed distributions, the normal scores were multiplied by specific \( g \) and \( h \) values so that the transformed scores followed the target non-normal distributions. Specifically, when \( g \) and \( h \) were nonzero,

\[ Y = \exp \left( hZ^2/2 \right) - \left[ \exp(gZ) - 1 \right] / g \]

(7)

where \( Y \) is the transformed score and \( Z \) is the original normal score. When \( g \) was zero,
\[ Y = Z \cdot \exp \left( hZ^2 / 2 \right) \]  

(8)

According to Algina et al., the target peaked distributions were manipulated at (1) \( Y_1 = 0 \) and \( Y_2 = 6 \) (i.e. \( g = 0 \) and \( h = 0.142 \)) and (2) \( Y_1 = 0 \) and \( Y_1 = 154.84 \) (i.e. \( g = 0 \) and \( h = 0.225 \)), and the target skewed distributions were fixed at (3) \( Y_1 = 2 \) and \( Y_1 = 6 \) (i.e. \( g = 0.76 \) and \( h = -0.098 \); an exponential distribution) and (4) \( Y_1 = 4.90 \) and \( Y_1 = 4,673.80 \) (i.e. \( g = 0.225 \) and \( h = 0.225 \)), which are common in social sciences research. Note that positively (or negatively) skewed distributions often have \( Y_1 > 0 \) (or \( Y_1 < 0 \), and short-tailed (or long-tailed; e.g. \( t \) distribution) distributions often have \( Y_2 < 0 \) (or \( Y_2 > 0 \)). For the mixed normal distribution, only 90% of the observations come from the normal distribution with mean 0 and SD 1 and 10% come from the normal distribution with mean 0 and SD 10. This distribution has \( Y_1 = 0 \) and \( Y_2 = 24.95 \).

**Factor 3: Number of DVs (\( v \); three levels).** Three numbers, 2, 5, and 8, were evaluated, which cover a range of values that seem to be practical in real-world research.

**Factor 4: Variance ratio (\( \pi \); three levels).** Variance ratio is defined as the ratio of the variance in Group 1 to the variance in Group 2 (Ruscio and Mullen, 2012). The ratio was fixed at 1, 4, and 0.25. The value of 1 means that the variances are homogenous for the two groups, and the values of 4 and 0.25 indicate violations of the homogeneity of covariance matrices assumption, which are commonly found in social sciences research.

**Factor 5: Correlations between DVs (\( R \); 2 \( \times \) 2 levels).** The DVs were expected to measure the same construct in MANOVA, and hence, they were manipulated to be correlated with one another in each group. Two levels of correlations, 0.50 and 0.80, were evaluated for the two groups, respectively. The value of 0.50 followed the design in Fouladi and Yockey (2002), which mimicked a moderate-to-large association between items. The value of 0.80 was included to examine the impact of extremely strong relationship. The manipulated levels in Factors 4 and 5 mimic the data conditions that meet or violate the assumption of homogeneity of covariance matrices.

**Factor 6: Total sample size (\( N \); two levels).** Two levels, 50 and 150, were simulated, thereby representing a small and moderate-large sample typically found in behavioral research.

**Factor 7: Base rate (\( b \); three levels).** In Ruscio and Mullen (2012), base rate is defined as the ratio of the sample size in Group 1 to the sample size in Group 2. Base rate was evaluated at the levels 0.25, 0.50, and 0.75. Thus, the samples sizes could be equal for the two groups, or it was three times larger in one sample than the other.

In sum, the factors were factorially combined to produce a design with \( 4 \times 6 \times 3 \times 2 \times 2 \times 2 \times 3 = 5,184 \) conditions. Each condition was replicated 10,000 times. The simulation was conducted in the R Project programming environment (R Core Team, 2014), and the code can be found in Supplemental Material.

**Evaluation criteria**

Two evaluation criteria were used. First, for each of the 5184 simulation conditions, percentage bias (bias) was computed to evaluate the average performance of an ES relative to its true value, that is, bias = \( \left( \frac{ES - \varphi}{\varphi} \right) \cdot 100\% \), where \( ES \) is the mean of the 10,000 simulated ESs, and \( \varphi \) is the true criterion ES value presented in standardized mean difference, eta squared, omega squared, or common-language metric. That is, as noted above, the true criterion values for \( \delta \) are 0.20, 0.50, and 0.80, and 1.50, \( \eta^2 \) are 0.0099, 0.0588, 0.1379, and 0.3600, and \( CL_w \) and \( A_w \) are 0.5793, 0.6915, 0.7881, and 0.9332. According to Li et al. (2011), a bias is considered reasonable if it is within \( \pm 10\% \). Second, to summarize the overall performance of an ES across 5184 conditions, mean absolute percentage error (MAPE) was computed, that is, \( \text{MAPE} = \sum_{i=1}^{5,184} |\text{bias} (i)| / 5,184 \). According to Brockwell and Davis (2002), A MAPE within 10% is regarded as a desirable fit.

**Results**

**Overall performance**

Comparing the four ESs, \( A_w \) performed the best. As shown in Figure 1, the biases ranged from \( -31.57\% \) to \( 52.35\% \), with a mean of \( -4.16\% \), which is within the criterion for a good fit \( (\pm 10\% \) . Of the 5184 conditions, 3548 (or 68.44%) resulted in a bias within \( \pm 10\% \) of the true ES. To summarize the overall performance across 5184 conditions, the MAPE was 8.20%, which is regarded as a desirable fit. The second reasonable ES in this study is \( CL_w \). The biases ranged from \( -42.19\% \) to \( 65.58\% \), with a mean of \( -7.67\% \), which is slightly beyond the criterion of \( \pm 10\% \). Of the 5184 conditions, 2666 (or 51.43%) produced a bias within \( \pm 10\% \). Regarding the overall performance, the MAPE was 12.02%, which is slightly beyond the criterion of \( \pm 10\% \).

By contrast, the conventional parametric ESs (\( \eta^2 \) and \( \omega^2 \)) were less than optimal. The biases of \( \eta^2 \) ranged from \( -94.13\% \) to \( 4884.73\% \), with a mean of \( 224.68\% \). Of the 5184 conditions, only 561 (or 10.82%) produced a bias within \( \pm 10\% \). The MAPE was 358.10%, showing unsatisfactory overall performance.
Note that the reason for a wide range of biases is in part due to the metrics of the true $\eta^2_\Lambda$ and $\omega^2_\Lambda$ in some conditions. For example, when $\delta=0.20$, $\Theta=normal$, $\nu=2$, $\pi=0.25$, $N=150$, $b=0.25$, and $R=(0.50, 0.50)$, the expected true $\eta^2_\Lambda$ is equal to 0.0099 and the expected true $\omega^2_\Lambda$ is equal to 0.0032. The means of the 10,000 simulated $\eta^2_\Lambda$ and $\omega^2_\Lambda$ values were found to be 0.0483 and 0.0417, respectively. Hence, the biases were very large (i.e. 388.20% and 1205.79%) because the simulated means were substantially larger than their true values. Despite the impacts of the metrics of the true $\eta^2_\Lambda$ and $\omega^2_\Lambda$ on the observed biases, $\eta^2_\Lambda$ and $\omega^2_\Lambda$ are still regarded as less accurate than the other ESs (i.e. $A_w$ and $CL_w$). The following sections discuss the effects of the manipulated factors on the ESs.

Effects of manipulated factors

Given that the correlations between DVs ($R$) did not show any effects on all the ESs, their impact is not discussed in this section. As shown in Table 1, first, when the total sample size ($N$) increased from 50 to 150, the accuracy of $\eta^2_\Lambda$, $A_w$, and $CL_w$ generally increased (biases decreased) because more data points were available for estimating the strength of the effect. The accuracy of $\omega^2_\Lambda$ only improved with a larger sample size when the true standardized difference ($\delta$) was larger than or equal to 0.50. When $\delta=0.20$, the biases changed from negative to positive with an increasing sample size. This was probably due to the effect of a small ES, which was less likely to be accurately detected or estimated. When $N$ was small, $\omega^2_\Lambda$ tended to underestimate the true value; when $N$ was large, $\omega^2_\Lambda$ tended to overestimate the true value.

Second, the variance ratio ($\pi$; i.e. variance ratio of the treatment group) did not show obvious impact on $A_w$ and $CL_w$. For $\eta^2_\Lambda$ and $\omega^2_\Lambda$, the biases were smaller when $\pi=1$ and $\delta \geq .50$. When $\pi \neq 1$, the biases depended on the level of $\delta$. When $\delta \geq .50$ and $\pi$ changed from 0.25 to 4.00, the biases changed from positive to negative. This was because the stronger group had a smaller (or larger) variance, and

---

**Figure 1.** Percentage biases of the generalized eta squared ($\eta^2_\Lambda$), generalized omega squared ($\omega^2_\Lambda$), non-parametric probability of superiority effect size ($A_w$), and parametric probability of superiority effect size ($CL_w$) across 5184 simulation conditions.
Table 1. Percentage biases of $\eta_s$, $\omega_s$, $A_w$, $C_w$ when the correlations between items were equal to (0.50, 0.50) for two groups.

| $\pi$ | $b$ | N | $\Theta = 0.5$ | $\Theta = 1$ | $\Theta = 2$ | $\Theta = 3$ | $\Theta = 4$ | $\Theta = 5$ | $\Theta = 6$ |
|--------|-----|----|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|        |     |    | $\eta_s$ | $\omega_s$ | $A_w$ | $C_w$ | $\eta_s$ | $\omega_s$ | $A_w$ | $C_w$ |
| 0.25   | 50  | 0.08 | 0.12 | 0.08 | 0.11 | 0.13 | 0.08 | 0.12 | 0.08 | 0.11 |
| 0.50   | 50  | 0.08 | 0.12 | 0.08 | 0.11 | 0.13 | 0.08 | 0.12 | 0.08 | 0.11 |
| 0.75   | 50  | 0.08 | 0.12 | 0.08 | 0.11 | 0.13 | 0.08 | 0.12 | 0.08 | 0.11 |
| 0.25   | 100 | 0.08 | 0.12 | 0.08 | 0.11 | 0.13 | 0.08 | 0.12 | 0.08 | 0.11 |
| 0.50   | 100 | 0.08 | 0.12 | 0.08 | 0.11 | 0.13 | 0.08 | 0.12 | 0.08 | 0.11 |
| 0.75   | 100 | 0.08 | 0.12 | 0.08 | 0.11 | 0.13 | 0.08 | 0.12 | 0.08 | 0.11 |

(Continued)
Table 1. (Continued)

| b   | N   | θ = 1 | v = 2 | ν = 3 | θ = 4 | v = 3 | θ = 5 | v = 5 |
|-----|-----|-------|-------|-------|-------|-------|-------|-------|
| 2.5 | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |
| 5.0 | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |
| 7.5 | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |
| 10  | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |
| 15  | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |
| 0.5 | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |
| 1.0 | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |
| 1.5 | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |
| 2.0 | 0.25| 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  | 0.38  |

(Continued)
| $\pi$ | $b$ | $N$ | $\Theta = 1$ | $\Theta = 2$ | $\Theta = 3$ | $\Theta = 4$ | $\Theta = 5$ | $\Theta = 6$ |
|------|-----|-----|-------------|-------------|-------------|-------------|-------------|-------------|
|      |     |     | $\eta_1^2$ | $\omega_1^2$ | $A_w$ | $CL_w$ | $\eta_1^2$ | $\omega_1^2$ | $A_w$ | $CL_w$ | $\eta_1^2$ | $\omega_1^2$ | $A_w$ | $CL_w$ | $\eta_1^2$ | $\omega_1^2$ | $A_w$ | $CL_w$ |
| 0.25 | 0.25 | 50  | 0.724 | 0.1120 | 0.04 | 0.008 | 5.21 | 8.05 | 0.02 | 0.00 | 6.85 | 10.62 | 0.01 | 0.00 | 6.58 | 10.17 | 0.00 | 0.01 | 6.99 | 10.81 | 0.06 | 0.17 |
|      | 1.00 | 50  | 0.908 | 0.602 | 0.05 | 0.002 | 2.94 | 4.53 | 0.01 | 0.00 | 3.51 | 5.41 | 0.02 | 0.00 | 3.32 | 5.11 | 0.00 | 0.00 | 3.43 | 5.28 | 0.00 | 0.01 |
| 0.50 | 0.25 | 150 | 1.87 | 2.11 | 0.02 | 0.004 | 0.46 | 0.52 | 0.00 | 0.00 | 1.30 | 1.47 | 0.11 | 0.11 | 1.04 | 1.18 | 0.13 | 0.13 | 0.94 | 1.00 | 0.17 | 0.17 |
|      | 1.00 | 50  | 1.69 | 2.60 | 0.03 | 0.012 | 1.27 | 1.95 | 0.00 | 0.00 | 1.34 | 2.06 | 0.05 | 0.04 | 1.19 | 1.83 | 0.03 | 0.03 | 1.15 | 1.76 | 0.03 | 0.08 |
| 0.75 | 0.25 | 150 | 0.81 | 0.92 | 0.05 | 0.003 | 0.36 | 0.29 | 0.00 | 0.00 | 0.37 | 0.41 | 0.02 | 0.01 | 0.17 | 0.19 | 0.00 | 0.00 | −0.30 | −0.34 | −0.03 | −0.20 |
|      | 1.00 | 50  | 2.85 | 4.38 | 0.05 | 0.002 | 2.32 | 3.56 | 0.00 | 0.00 | 2.56 | 3.95 | 0.03 | 0.03 | 2.44 | 3.75 | 0.05 | 0.09 | 2.44 | 3.75 | 0.02 | 0.04 |
| 1.00 | 0.25 | 150 | 1.13 | 1.28 | 0.01 | 0.006 | 0.15 | 0.17 | −0.08 | 0.00 | 0.74 | 0.83 | 0.04 | 0.12 | 0.56 | 0.63 | 0.05 | 0.15 | 0.56 | 0.63 | 0.05 | 0.15 |
|      | 1.00 | 50  | 3.20 | 4.92 | 0.08 | 0.006 | 2.47 | 3.80 | 0.01 | 0.00 | 2.82 | 4.35 | 0.03 | 0.01 | 2.66 | 4.09 | 0.04 | 0.01 | 3.03 | 4.61 | 0.06 | 0.04 |
| 0.50 | 0.25 | 150 | 1.50 | 1.70 | 0.02 | 0.005 | 0.22 | 0.25 | −0.08 | 0.00 | 1.00 | 1.13 | −0.21 | 0.14 | 0.76 | 0.86 | 0.14 | 0.14 | 1.14 | 1.29 | 0.01 | 0.01 |
|      | 1.00 | 50  | 2.83 | 4.36 | 0.14 | 0.017 | 2.30 | 3.53 | 0.06 | 0.00 | 2.54 | 3.91 | 0.12 | 0.13 | 2.41 | 3.71 | 0.10 | 0.10 | 2.69 | 4.14 | 0.09 | 0.10 |
| 0.75 | 0.25 | 150 | 1.14 | 1.29 | 0.04 | 0.005 | 0.15 | 0.17 | −0.06 | 0.00 | 0.74 | 0.84 | 0.11 | 0.12 | 0.56 | 0.63 | 0.01 | 0.01 | 0.85 | 0.96 | 0.02 | 0.02 |
|      | 1.00 | 50  | 2.05 | 3.08 | 0.06 | 0.006 | 0.74 | 1.13 | −0.01 | 0.00 | 0.51 | 0.78 | 0.02 | 0.07 | 10.11 | 10.70 | 0.04 | 0.05 | 10.55 | 8.85 | 0.01 | 0.07 |
| 1.00 | 0.25 | 150 | 0.63 | 0.96 | 0.04 | 0.005 | 0.74 | 1.13 | −0.01 | 0.00 | 0.51 | 0.78 | 0.02 | 0.07 | 10.11 | 10.70 | 0.04 | 0.05 | 10.55 | 8.85 | 0.01 | 0.07 |
|      | 1.00 | 50  | 2.68 | 4.12 | 0.09 | 0.006 | 0.24 | 0.44 | 0.00 | 0.00 | 2.54 | 3.92 | 0.08 | 0.08 | 2.47 | 3.80 | 0.07 | 0.16 | 2.45 | 3.76 | 0.07 | 0.07 |
| 0.50 | 0.25 | 150 | 0.49 | 0.55 | 0.04 | 0.006 | 0.09 | 0.08 | 0.00 | 0.00 | 0.32 | 0.37 | −0.06 | 0.15 | 0.25 | 0.28 | 0.07 | 0.16 | 1.54 | 1.75 | 0.03 | 0.03 |
|      | 1.00 | 50  | 6.50 | 10.05 | 0.26 | 0.037 | 4.73 | 7.30 | 0.17 | 0.28 | 6.23 | 9.63 | 0.25 | 0.36 | 6.00 | 9.28 | 0.24 | 0.36 | 5.94 | 9.18 | 0.21 | 0.37 |
| 0.75 | 0.25 | 150 | 1.77 | 2.00 | 0.01 | 0.002 | 0.94 | 1.07 | −0.04 | 0.00 | 1.61 | 1.82 | 0.00 | 0.00 | 1.51 | 1.70 | 0.00 | 0.01 | 2.56 | 2.90 | 0.00 | 0.15 |

$\pi$ is the variance ratio, $b$ is the base rate, $N$ is the total sample size, $\nu$ is the number of dependent variables, $\delta$ is the true effect size expressed as the standardized mean difference. $\Theta$ indicates six distributions: 1 = normal, 2 = mixed normal, 3 = peaked ($\Gamma_1 = 0$ and $\Gamma_2 = 6$), 4 = peaked ($\Gamma_1 = 6$ and $\Gamma_2 = 15.84$), 5 = skewed ($\Gamma_1 = 2$ and $\Gamma_2 = 6$), and 6 = skewed ($4.90$ and $\Gamma_2 = 4.67380$), where $\Gamma_1$ refers to skewness and $\Gamma_2$ refers to kurtosis. $\eta^2_1$ is the generalized eta squared, $\omega^2_1$ is the non-parametric probability of superiority effect size, and $CL_w$ is the parametric probability of superiority effect size.
hence, the estimated sum of squares became over-precise (or under-precise) so that a stronger ES resulted for $\pi = .25$ (or $\pi = .4$). When $\delta \leq .2$, there was no obvious pattern of relationships between $\pi$ and biases.

Third, the base rate ($b$; i.e. proportion of the sample size of the control group) did not follow a clear pattern of relationship with $\omega^2_\lambda$, $A_w$ and $CL_w$. For $\eta^2_\lambda$, the effects were more complicated. When $\delta \leq .50$ and $\pi = 1$, $b$ did not show much impact on $\eta^2_\lambda$. When $b$ changed from 0.25 to 0.75 and $\pi = .25$ bias generally decreased; but when $\pi = 4$, biases generally increased. There was no clear explanation for this pattern, but we are sure that the accuracy of $\eta^2_\lambda$ fluctuated substantially with varying levels of the base rate and variance ratio, and thus, it was not robust to these factors.

Fourth, when the true ES ($\delta$) increased, the biases of $\eta^2_\lambda$ and $\omega^2_\lambda$ generally decreased due to the magnitude of the effect being more accurately detected when it is large. The other ESs ($A_w$ and $CL_w$) were generally accurate regardless of the magnitude of $\delta$.

Fifth, when the number of DVs ($v$) increased, the accuracy of $\eta^2_\lambda$ and $\omega^2_\lambda$ generally decreased. This was probably due to the more complicated mathematical procedure required to obtain the $a$ weights (equation (4)) that maximized the weighted difference, given that other factors were held constant. On the other hand, $A_w$ and $CL_w$ were generally robust to the increased number of DVs.

Sixth, it came as a surprise that the performance of $\eta^2_\lambda$, $\omega^2_\lambda$, and $CL_w$ did not differ across the six distributions. For $\eta^2_\lambda$, this was because the sum-of-squares based Wilks Lambda ($A$; equation (1)) appeared to be robust to any shape of the distribution, if both groups shared the same distribution. That is, the sum of squares measures the degree to which the scores surround the mean. When the two groups shared the same distribution, the sum of squares would measure the variability of the scores on the same metric, and hence, the Wilks Lambda as well as $\eta^2_\lambda$ and $\omega^2_\lambda$ were unaffected by non-normal distributions. On the other hand, if the distributions are different for the two groups (e.g. normal and skewed), $\eta^2_\lambda$ and $\omega^2_\lambda$ may become inaccurate, but the present study did not examine this factor. $CL_w$ was also found to be robust to any shape of distribution. It appeared that the parametric mean difference estimator (i.e. $\left( w_p - w_q \right) / s_w$; equation (6)) was also robust to non-normal distributions, when the two groups shared the same non-normal distributions. For the last ES ($A_w$), its mathematical equation did not rely on any parametric assumption, and hence, it was robust to any of the six distributions as expected.

**Conclusion and discussion**

This study proposes and develops a non-parametric ES ($A_w$) for MANOVA. The results of a Monte-Carlo simulation showed that $A_w$ was accurate across the simulated conditions and robust to violations of the two key assumptions (multivariate normality and homogeneity of covariance matrices).

It also outperformed the two conventional parametric ESs (i.e. $\eta^2_\lambda$ and $\omega^2_\lambda$) and its parametric counterpart $CL_w$. Hence, researchers and practitioners are encouraged to report $A_w$ for ES evaluation in a MANOVA scenario, especially when the key assumptions are violated.

The proposed $A_w$ is important for researchers in behavioral and social sciences research because evaluating the conventional ESs ($\eta^2_\lambda$ and $\omega^2_\lambda$) could be misleading in the existing literature. Our findings showed that $\eta^2_\lambda$ could be 4297% larger than its true ES when the homogeneity of variance assumption was violated (i.e. variance ratio = 0.25), the sample size ratio of the control group to the treatment group was 0.25, the number of DVs was 8, the sample size was 50, and the true ES was small (0.20). By the same token, the conventional $\omega^2_\lambda$ could be 4088% below its true value under the same conditions. These findings align with a recent claim by Hoekstra et al. (2012) who strongly encouraged researchers to check the assumptions for conventional analyses. The authors found that only 33.3% and 25% of researchers in psychology have correctly checked the assumptions of homogeneity of variances and normality, respectively, for conventional and widely-used analyses such as t-test, ANOVA, and regression. The majority of the participants in Hoekstra et al.’s study stated that they were either unfamiliar with these assumptions or did not understand how to check these assumptions. Moreover, Ruscio and Roche (2012) reviewed 455 studies published in leading psychology journals and found that the reported variance ratios often fail to meet the requirement, thereby posing a threat or risk of inflated Type 1 error rates when using conventional, parametric statistical methods. We hope that as researchers in psychology start to embrace the recommended practice of reporting ESs, they will be more aware of the importance of these assumptions and will consider adopting non-parametric alternatives where appropriate.

Considering the current lack of awareness about the importance of parametric assumptions for conventional ESs, and the prevalence of research situations in which the assumptions are violated, this study offers a non-parametric alternative ($A_w$) for MANOVA that does not depend on these assumptions.

The implications of this study can be generalized to researchers and practitioners in a wide range of other disciplines, both in social and natural sciences, who often use MANOVA. For example, clinical trials researchers often examine the difference between a treatment group and a placebo group in a number of health-related criterion measures (e.g. body mass index, blood pressure). On some occasions, their data may violate the assumptions for the traditional ESs, and hence, the proposed $A_w$ can provide a more trustworthy measure for evaluating the difference between the two groups. Biological researchers are often interested in comparing the difference between an experimental group and a control group in a lab setting (e.g. effects of room temperature and absolute zero degree on cellular motility and signaling), and they could also report $A_w$ for this scenario.
Limitations and directions for future research

A first area of ongoing research lies in examining the effects of different distributions for the two groups of observations. In this study, the two groups of scores were assumed to follow the same (either normal or non-normal) distributions. Future research should include a simulation study to examine the effects of unbalanced distributions (e.g. normal vs skewed) on the ESs in MANOVA.

A second area of research involves generalization of the proposed $A_w$ to the one-way MANOVA with more than two independent samples as well as to the multi-way MANOVA that involves multiple IVs (e.g. factorial and mixed designs). These more general or complicated types of MANOVA are also popular in psychology research. This study lays foundation for the non-parametric ES in a simpler MANOVA. Ruscio and Gera (2013) have recently provided the extensions of $A_w$ to one-way ANOVA. Additional research can derive mathematical equations for one-way MANOVA based on Ruscio and Gera’s study and provide empirical evidence for this statistic based on a simulation study.

In addition to the reporting of ES, the new statistical practices suggest that researchers report the CIs surrounding a reported ES. Therefore, a third area for future research is examination of the sampling distribution or confidence intervals (CIs) surrounding the proposed $A_w$. Ruscio and Mullen (2012) found that the bootstrap CIs constructed for the non-parametric $A$ in ANOVA were accurate. Further research can also examine the use of the bootstrap procedure for the CIs surrounding the proposed $A_w$ in MANOVA.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the University Research Grant Program (#43869) at the University of Manitoba.

ORCID iDs

Johnson Ching-Hong Li https://orcid.org/0000-0002-2553-117X
Rory Michael Waisman https://orcid.org/0000-0003-2608-169X

Supplemental material

Supplemental material for this article is available online.

Note

1. Steyn and Ellis (2009) also discussed a derivative to $\eta^2_k$—generalized partial eta squared $\eta^2_{k,p}$—that measures the unique contribution of an IV with other IVs being partial out. However, this study focuses on MANOVA with only one IV, and hence, $\eta^2_{k,p} = \eta^2_k$. Moreover, Steyn and Ellis discussed the Rao $F$ statistic, which is the ratio of explained variance over the unexplained variance. However, it is not applicable when there are only two levels in the IV and two DVs, and hence, it is not evaluated in this study.

References

Algina J, Keselman HJ and Penfield RD (2005) An alternative to Cohen’s standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. Psychological Methods 10: 317–328.

Algina J, Oshima TC and Tang KL (1991) Robustness of Yao’s, James’ and Janssen’s tests under variance-covariance heteroscedasticity and nonnormality. Journal of Educational Statistics 16: 125–139.

American Psychological Association (2010) Publication Manual of the American Psychological Association, 6th edn. Washington, DC: Author.

Brockwell PJ and Davis RA (2002) Introduction to Time Series and Forecasting, New York: Springer.

Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole DA, Maxwell SE, Arvey R, et al. (1994) How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. Psychological Bulletin 115: 465–474.

Everitt BS (1979) A Monte Carlo investigation of the robustness of Hotelling’s one- and two-sample $T^2$ tests. Journal of the American Statistical Association 74: 48–51.

Finch WH (2016) Comparison of multivariate means across groups with ordinal dependent variables: A Monte Carlo simulation study. Frontiers in Applied Mathematics and Statistics 2(2): 1–11.

Fouladi RT and Yockey RD (2002) Type I error control of two-group multivariate tests on means under conditions of heterogeneous correlation structure and varied multivariate distributions. Communications in Statistics – Simulation and Computation 31: 375–400.

Govindarajulu Z (1967) Two-sided confidence limits for $P(X < Y)$ for normal samples of $X$ and $Y$. The Indian Journal of Statistics Series B 29(1/2): 35–40.

Hoeckstra R, Kiers HA and Johnson A (2012) Are assumptions of well-known statistical techniques checked, and why (not)? Frontiers in Psychology 3: 137.

Hopkins JW and Clay PPF (1963) Some empirical distributions of bivariate $T^2$ and homoscedasticity criterion M under unequal variance and leptokurtosis. Psychological Methods 9: 183–197.

Keselman HJ and Lix LM (1997) Analysing multivariate repeated measures designs when covariance matrices are heterogeneous. British Journal of Mathematical and Statistical Psychology 50: 319–338.

Li JCH, Chan W and Cui Y (2011) Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. British Journal of Mathematical and Statistical Psychology 64: 367–387.

Mann HB and Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics 18(1): 50–60.
Mathews J and Mathews R (2007) Successful Scientific Writing: A Step-by-Step Guide for the Biological and Medical Sciences. Cambridge, MA: Cambridge University Press.

McGraw KO and Wong SP (1992) A common language effect size statistic. Psychological Bulletin 111: 361–365.

Mordkoff JT (2019) A simple method for removing bias from a popular measure of standardized effect size: Adjusted partial eta squared. Advances in Methods and Practices in Psychological Science 2: 228–232.

Murphy KR (1997) Editorial. Journal of Applied Psychology 82: 3–5.

Olejnik S and Huberty CJ (2006) Applied MANOVA and Discriminant Analysis. Ukraine: Wiley.

R Core Team (2014) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org/ (accessed 21 February 2015).

Ruscio J (2008) A probability-based measure of effect size: Robustness to base rates and other factors. Psychological Methods 13(1): 19–30.

Ruscio J and Gera BL (2013) Generalizations and extensions of the probability of superiority effect size estimator. Multivariate Behavioral Research 48: 208–219.

Ruscio J and 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.

Ruscio J and Roche B (2012) Variance heterogeneity in published psychological research: A review and a new index. Methodology 8: 1–11.

Steyn HS Jr and Ellis SM (2009) Estimating an effect size in one-way multivariate analysis of variance (MANOVA). Multivariate Behavioral Research 44: 106–129.

Tang KL and Algina J (1993) Performance of four multivariate tests under variance-covariance heteroscedasticity. Multivariate Behavioral Research 28(4): 391–405.

Thompson B (1994) Guidelines for authors. Educational and Psychological Measurement 54(4): 837–847.

Trafimow D and Marks M (2015) Editorial. Basic and Applied Social Psychology 37(1): 1–2.

Troncoso Skidmore S and Thompson B (2013) Bias and precision of some classical ANOVA effect sizes when assumptions are violated. Behavior Research Methods 45: 536–546.

Vargha A and Delaney HD (2000) A critique and improvement of the CL common language effect size statistic of McGraw and Wong. Journal of Educational and Behavioral Statistics 25: 101–132.

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1: 80–83.

Wolfe DA and Hogg RV (1971) On constructing statistics and reporting data. The American Statistician 25(4): 27–30.

Author biographies

Dr. Johnson Ching-Hong Li is an Associate Professor in the Department of Psychology at the University of Manitoba. His research interests include effect size estimates, statistical corrections for study artifacts, evaluation of psychological scales, meta-analysis, robust statistics, probability-based statistics, structural equation modeling, cognitive diagnostic modeling, and reliability assessments.

Mr. Marcello Nesca is a Master’s student in community health sciences with a thesis focus on data quality methods for text data. Currently, he is employed at the Manitoba Centre for Health Policy and is working on projects involving using natural language processing and machine learning methods on electronic medical records.

Mr. Rory Michael Waisman is a Ph.D. candidate in the Department of Marketing, Business Economics, and Law at the University of Alberta. He is broadly interested in how consumers construct judgments, decisions, and preferences. His research has three primary streams: evaluation of consumption objects and experiences, financial decision making, and persistence of the behavioral effects of nudges.

Mr. Yongtian Cheng is a Ph.D. student in the Division of Psychology and Language Sciences at University College London (UCL). He received his B.A. (Hons.) and M.A. from the University of Manitoba. His major research area is quantitative psychology, with an emphasis on replication crisis, effect size measures, and Monte Carlo simulation.

Dr. Virginia Man Chung Tze is an Assistant Professor in the Department of Educational Administration, Foundations and Psychology at the University of Manitoba. Her research focuses on social-emotional learning, including emotional management, self-regulation, and cultural diversity.