Interval Estimation for the Difference between Correlated Proportions: An Empirical Investigation of Accuracy and Precision of Three Methods

Jeanine L. Romano[1], Eun Sook Kim[1], Patricia Rodriguez de Gil[1], Thanh Pham[1],Pei-Chen Wu[1], Diep Nguyen[1], Jeffrey Kromrey[1]
[1]University of South Florida, 4202 E. Fowler Ave. 105, Tampa, FL 33620

Abstract
Three methods for computing confidence intervals (CI) around differences in correlated proportions (Wald CI, adjusted Wald CI, and a likelihood-based CI method proposed by Tango (1998)) were investigated to determine which of these methods produces the most accurate and precise CI estimates. Two dichotomous outcomes measured from dependent samples were simulated. The factors manipulated in the simulation study included overall sample size (10, 50, 100, 500, 1000), direction and strength of the relationship between the two proportions (±.40, ±.30, ±.20, ±10, 0), and the population difference in marginal proportions (±.3, ±.25, ±.10, ±.05, 0). For each sample generated (i.e., 100,000 replications), each of the three proposed CI methods was calculated. The adjusted Wald CI provided the best coverage across the conditions investigated and is easier to calculate than the Tango intervals. In addition, both the original Wald CI and the Tango CI produced substantial undercoverage in some small sample conditions.

Key words: Interval estimates, Correlated Proportions, Dichotomous Outcomes.

1.Background

The analysis of correlated or paired data (e.g., repeated measures, pretest-posttest, rater-agreement) is frequently encountered in applied research; that is, research designs in which a sample of *n* units provide two dichotomous responses that can be summarized in a 2 x 2 contingency table. One example is the results of two diagnostic tests of dyslexia (test A and test B), administered to a sample of pre-school children and summarized in Table 1 as cell frequencies of the cases in which diagnosis is correct in both tests (correct-correct), correct in test A but incorrect in test B (correct-incorrect), incorrect in test A but correct in test B (incorrect-correct), and incorrect in both tests (incorrect-incorrect).

Table 1: Contingency Table of Dyslexia Diagnosis from Test A and B

		Test B		
		Correct	Incorrect	Total
Test A	Correct	a (π_{11})	b (π_{12})	a + b (π_{1+})
	Incorrect	c (π_{21})	d (π_{22})	c + d (π_{2+})
	Total	a + c (π_{+1})	b + d (π_{+2})	*n* (1)

The difference in correlated or dependent proportions is often of interest in studies such as pretest-posttest designs, matched-pair designs, and rater-agreement designs (Bonett & Price, 2011). Currently, McNemar's test (1947) is commonly conducted to test the equivalence of two correlated proportions. In this example, the null hypothesis is that the population proportion of correct diagnosis on dyslexia using Test A (π_{1+} = [a + b]/n) equals the population proportion of correct diagnosis using Test B (π_{+1} = [a + c]/n). Testing the null hypothesis $\pi_{1+} - \pi_{+1} = 0$ is equivalent to testing $\pi_{12} - \pi_{21} = 0$ because π_{11} is common to both π_{1+} and π_{+1}. Given the null hypothesis McNemar's test statistic (*Q*) is computed as

$$Q = \frac{(b - c)^2}{b + c}.$$

Under the null hypothesis, the *Q* statistic follows an asymptotic chi-square distribution with one degree of freedom when b + c is greater than 10 (McNemar, 1947). A 100(1 – α)% Wald confidence interval for the difference in the population proportions ($\pi_{12} - \pi_{21}$) can be estimated as

$$\hat{\pi}_{12} - \hat{\pi}_{21} \pm z_{\frac{\alpha}{2}}\sqrt{[\hat{\pi}_{12} + \hat{\pi}_{21} - (\hat{\pi}_{12} - \hat{\pi}_{21})^2]/n}$$

where $z_{\alpha/2}$ is a critical value at α/2 from the standard normal distribution. However, when Newcombe (1998) examined the performance of existing methods to compute the confidence intervals for the difference in correlated proportions, the Wald confidence interval showed inadequate performance.

Recently, Bonett and Price (2011) proposed an alternative CI by making an adjustment to the Wald interval.

$$\hat{\pi}_{12} - \hat{\pi}_{21} \pm z_{\frac{\alpha}{2}}\sqrt{[\hat{\pi}_{12} + \hat{\pi}_{21} - (\hat{\pi}_{12} - \hat{\pi}_{21})^2]/(n + 2)}$$

where each cell proportion is computed by adding one to the cell frequency and two to the total *n*. For example, $\hat{\pi}_{12} = (b + 1)/(n + 2)$ and $\hat{\pi}_{21} = (c + 1)/(n + 2)$. Bonett and Price (2011) reported that the adjusted Wald interval performs as well as an approximate CI proposed by Tango (1998) but the computation is simpler than Tango's.

The confidence interval for the difference in two correlated proportions ($\lambda = \pi_{1+} - \pi_{+1} = \pi_{12} - \pi_{21}$) developed by Tango is estimated by solving the following two equations iteratively until the change in estimation is infinitesimal below the predetermined cutoff.

$$\frac{b - c - n\lambda}{\sqrt{n(2\hat{\pi}_{21} + \lambda(1 - \lambda))}} = \pm z_{\alpha/2},$$

and $\hat{\pi}_{21}$ is estimated as

$$\hat{\pi}_{21} = \frac{\sqrt{(B^2 - 4AC)} - B}{2A}$$

where A = 2n, B = -b – c + (2n – b + c)λ, and C = -cλ(1 – λ). Although the computational procedures are more complex than Wald and adjusted Wald intervals, the upper and

lower limits are easily found through the secant method with empirically good coverage probabilities (Tango, 1999) and can be applied to small samples with off-diagonal zero cells (Tango, 1998). A SAS macro for the computation of the Tango interval, as well as the two Wald intervals, was provided by Wu et al. (2013).

2. Purpose

The purpose of this study was to examine three different methods for computing confidence intervals around differences in correlated proportions: Wald's method, the adjusted Wald method proposed by Bonett and Price (2011) and the likelihood-based method proposed by Tango (1998). The accuracy and precision of the confidence interval estimation methods were investigated using Monte Carlo methods by simulating two dichotomous outcomes measured from dependent samples.

3. Method

In this simulation study the following design factors and conditions were examined: (a) sample size (N = 10, 50, 100, 500, and 1000), (b) magnitude and direction of population correlation between the two proportions (φ = -.40 to +.40, in increments of .10), and (c) difference between population proportions (Δ = -.30, -.25, -.10, -.05, .00, .05, .10, .25, and .30). These factors in the Monte Carlo study were completely crossed, yielding 405 conditions. For each condition, 100,000 replications were conducted and in each simulated sample the three confidence interval methods were applied to provide estimates of 90%, 95% and 99% intervals. The use of 100,000 estimates provides adequate precision for the investigation of confidence interval coverage and width. For example, 100,000 replications provide a maximum 95% confidence interval width around an observed proportion that is ± .0031 (Robey & Barcikowski, 1992).

The data for the simulation were generated using uniform random numbers on the zero to one interval (the SAS RANUNI function). The values of the random numbers were used to assign observations to cells in the contingency table. For example, to simulate data with Δ = .25 and φ = .00, observations with random numbers between 0 and .21 were assigned to the cell in the first row and first column, those with random numbers between .21 and .60 were assigned to the cell in the first row and second column, and so on. This procedure yields 2 X 2 tables in which the expected cell proportions are illustrated in Table 2.

For the population presented in Table 2, the difference in proportions (Δ) is .39 - .14 = .25, and the correlation is given by

$$\varphi = \frac{.26 - (.39 + .26)(.14 + .26)}{\sqrt{(.21 + .39)(.14 + .26)(.21 + .14)(.39 + .26)}} = 0$$

The simulation was conducted using SAS/IML (SAS, 2008). The data simulation program was checked by examining the matrices produced at each stage of data generation. Primary outcomes examined in this Monte Carlo study were the confidence interval coverage probabilities and confidence interval widths over the 100,000 replications.

Table 2: Expected Cell Proportions with $\Delta = .25$ and $\varphi = .00$

		X2	
		1	2
X1	1	.21	.39
	2	.14	.26

4. Results

The outcomes from the simulation were analyzed by first examining box plots to describe the distributions of each outcome across all simulation conditions. In addition, the results of the study were evaluated by using analysis of variance to compute the effect size η^2 associated with each of the research design factors and their first-order interactions. To save space, the results from the 95% confidence intervals are presented. Results for 90% and 99% intervals are available from the authors.

4.1 Interval Coverage

The overall distributions of interval coverage results at 95% are represented in Figure 1. The results suggest that adjusted Wald method provides the best interval coverage overall among the three methods. The original Wald method presents lower coverage than the expected confidence level in the majority of conditions. The adjusted Wald and Tango methods provide higher coverage for most of the sample conditions. However, the interval coverage of the adjusted Wald method is closer to the expected confidence level than that of the Tango method. Similar results were seen at the 90% and 99% confidence levels.

In the analysis of the simulation design factors associated with variability in the interval coverage, estimation method, the population delta, the interaction between method and sample size, as well as the interaction between method and population delta resulted in large effects ($\eta^2 = .30$, $\eta^2 = .07$, $\eta^2 = .25$ and $\eta^2 = .11$, respectively). Similar results were seen for the 90% and 99% confidence interval estimates.

The mean coverage estimates associated with the interaction between method and sample size are presented in Figure 2 and Table 3. Clearly the Tango method provides the best coverage at the small sample size but provides slight over coverage with large sample sizes. The original Wald method evidences a large degree of under coverage at the smallest sample size and gets closer to the nominal confidence interval coverage when sample size increases. The adjusted Wald method presents slight over coverage at the smallest sample size but the difference is smaller when sample size increases. The means of interval coverage by sample size for each estimation method are summarized in Table 3.

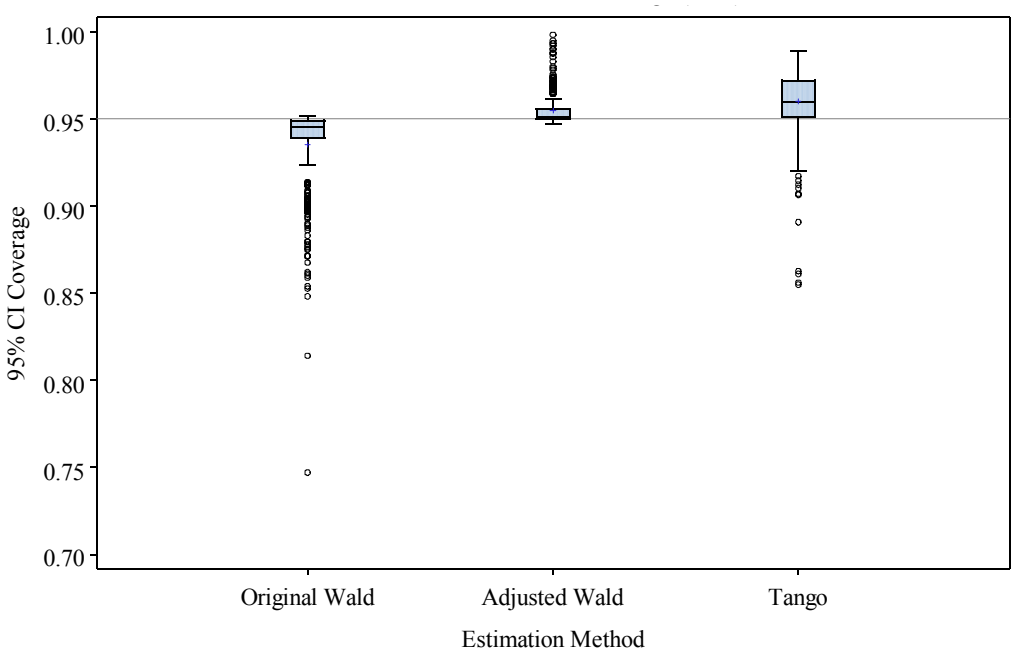

Figure 1: The distributions of interval coverage at 95% confidence level.

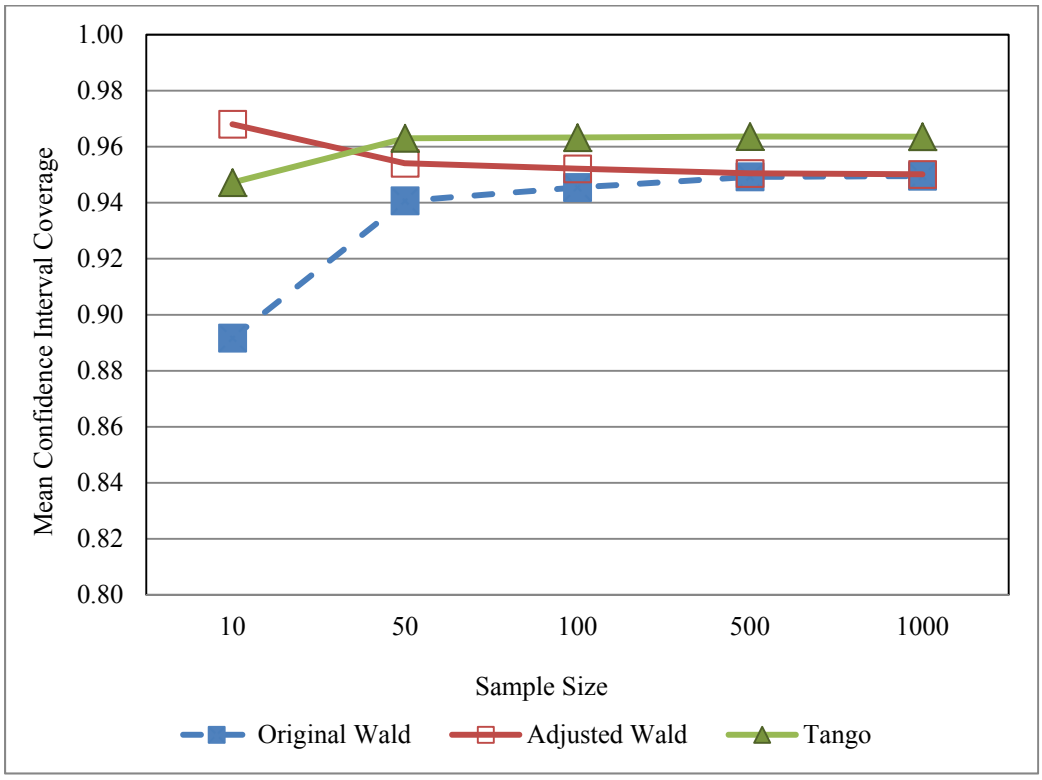

Figure 2: The interaction between methods and sample size on coverage level at 95% confidence

Table 3: Mean Interval Coverage by Sample Size for Three Methods

	Method		
Sample size	Original Wald	Adjusted Wald	Tango
10	0.89	0.97	0.95
50	0.94	0.95	0.96
100	0.95	0.95	0.96
500	0.95	0.95	0.96
1000	0.95	0.95	0.96

Note. Estimates are based on 100,000 samples of each condition.

Table 4 displays the mean interval coverage estimates by population delta for each estimation method. The interaction between method and population delta on coverage is also shown in Figure 4. Based on this figure, the adjusted Wald provides mean coverage close to the nominal 95% percent confidence at all level of population delta while the original Wald method results in under coverage at all population delta levels. The Tango method provides near nominal coverage only at population delta values close to zero. Notably, the Tango method provides greater over coverage when delta gets farther from zero.

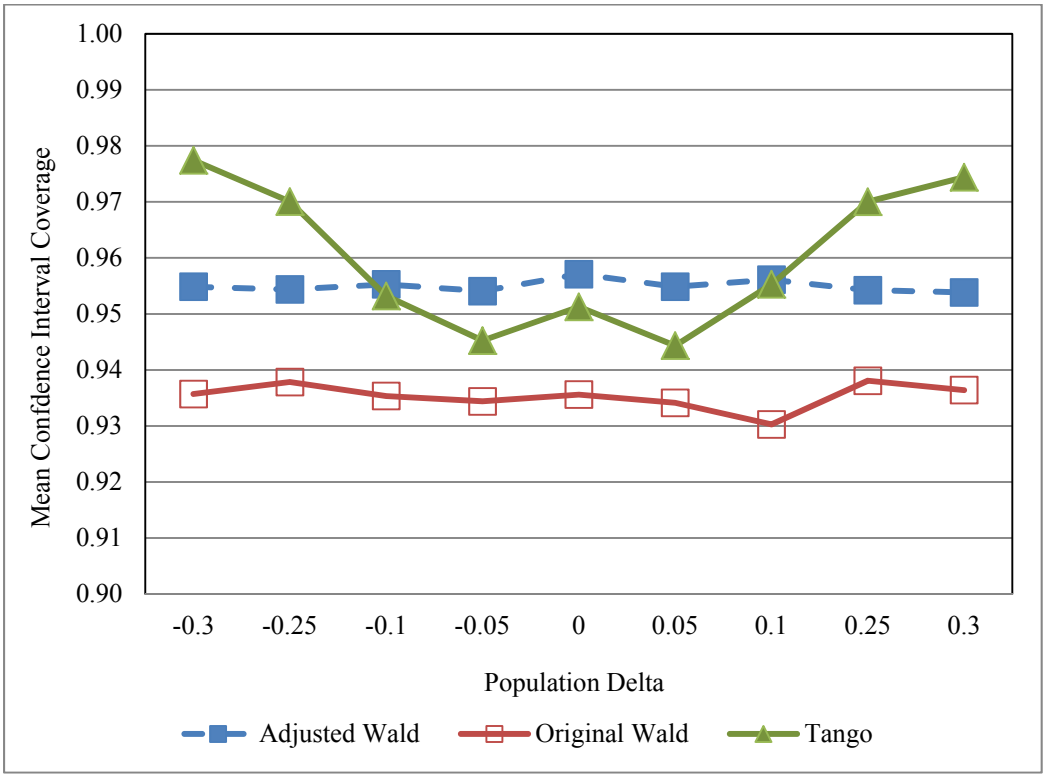

Figure 4: The interaction between method and population delta for coverage level at 95% confidence

Table 4: Mean Interval Coverage by Population Delta for Three Methods

	Method		
Population delta	Original Wald	Adjusted Wald	Tango
-0.30	0.94	0.95	0.98
-0.25	0.94	0.95	0.97
-0.10	0.94	0.96	0.95
-0.05	0.93	0.95	0.95
0.00	0.94	0.96	0.95
0.05	0.93	0.95	0.94
0.10	0.93	0.96	0.96
0.25	0.94	0.95	0.97
0.30	0.94	0.95	0.97

Note. Estimates are based on 100,000 samples of each condition.

4.1.1 Evaluation of the adequacy of the estimated intervals

Bradley's (1978) liberal criterion provides a guideline for evaluating the robustness of a hypothesis testing method. For a given nominal alpha level, Type I error rates between $\alpha - 0.5\alpha$ and $\alpha + 0.5\alpha$ are conceived as robust at the most liberal end. We adopted Bradley's liberal criterion to evaluate the adequacy of the estimated intervals. When the proportion of intervals that contain the population parameter falls between $1 - 1.5\alpha$ and $1 - 0.5\alpha$ (for example, between .925 and .975 for a 95% confidence interval), a method is considered adequate with respect to the interval coverage.

First, the overall proportions of simulation conditions that meet Bradley's criterion were examined. Adjusted Wald intervals met the Bradley's criterion most frequently (.96) followed by original Wald (.80) and Tango (.76) intervals. To examine the adequacy of the estimated intervals with small samples, the proportions meeting the Bradley's criterion were investigated by sample size (see Table 5). When sample size is small (N = 10), the performances of all three intervals deteriorated considerably with respect to Bradley's criterion. However, the proportions meeting Bradley's criterion for the adjusted Wald intervals were relatively high compared to the other methods. On the other hand, the impact of small sample size was the most salient with the original Wald intervals. For example, when N = 10, the proportion meeting the Bradley' criterion is .01 at α = .05. Based on this observation, the original Wald method is not recommended with very small samples.

Table 5: Proportion of Conditions Meeting the Bradley's Liberal Criterion by Sample Size

N	Wald	Adjusted Wald	Tango
10	.01	.80	.74
50	1.00	1.00	.78
100	1.00	1.00	.77
500	1.00	1.00	.75
1000	1.00	1.00	.77

Note. Estimates are based on 100,000 samples of each condition.

Compared to the two Wald methods, the Tango intervals were frequently inadequate in terms of Bradley's criterion. That is, the empirical estimates of the coverage provided by the Tango intervals often missed the Bradley's liberal criterion showing typically over coverage. In addition, the Tango method showed a large variability in the Bradley proportions across the levels of simulation conditions. Eta-squared was analyzed to identify primary simulation factors explaining the variability of the Bradley proportions in the Tango intervals. The direction and strength of relationship between the two proportions (φ), the population difference in marginal proportions (Δ), and the interaction between them are highly associated with the Bradley proportions (η^2 = .19, .23, and .25, respectively). When the difference in marginal proportions (Δ) is close to zero, the proportions meeting the Bradley's criterion are 1.0 or close to 1.0 (see Table 7). When Δ = 0, the Tango intervals met Bradley's criterion all the time across the levels of φ. In contrast, as Δ departs from 0, the adequacy of Tango intervals depends on both direction and strength of relationship of two proportions (φ). When φ < 0, the Tango intervals showed high levels of adequacy in the interval coverage. However, when φ ≥ 0, that is, when the difference in proportions was reversed, the Bradley's criterion was less frequently met as Δ departs from zero.

Table 7: Proportion of Conditions Meeting Bradley's Criterion for the Tango Intervals by φ and Δ

	φ								
Δ	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4
-0.30	1.0	1.0	.40	.00	.20	.20	.20	.00	.00
-0.25	1.0	1.0	1.0	.20	1.0	1.0	.80	.20	.00
-0.10	1.0	1.0	1.0	.80	1.0	1.0	1.0	.80	.80
-0.05	1.0	1.0	1.0	1.0	1.0	.80	.80	.80	.80
0.00	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0.05	1.0	1.0	1.0	1.0	.80	.80	.80	.80	.80
0.10	1.0	1.0	1.0	.80	1.0	1.0	.20	.80	.80
0.25	1.0	1.0	1.0	1.0	1.0	1.0	.20	.20	.20
0.30	1.0	1.0	1.0	1.0	.20	.20	.20	.00	.00

Note. φ = the direction and strength of relationship of two proportions, Δ = the population difference in marginal proportions.

4.2 Interval Width

Figure 5 displays the distributions of interval widths at 95% confidence. The results indicate that these distributions are almost the same across all three estimation methods. The distributions of interval widths are also very similar across three estimation methods at 90% and 99% confidence.

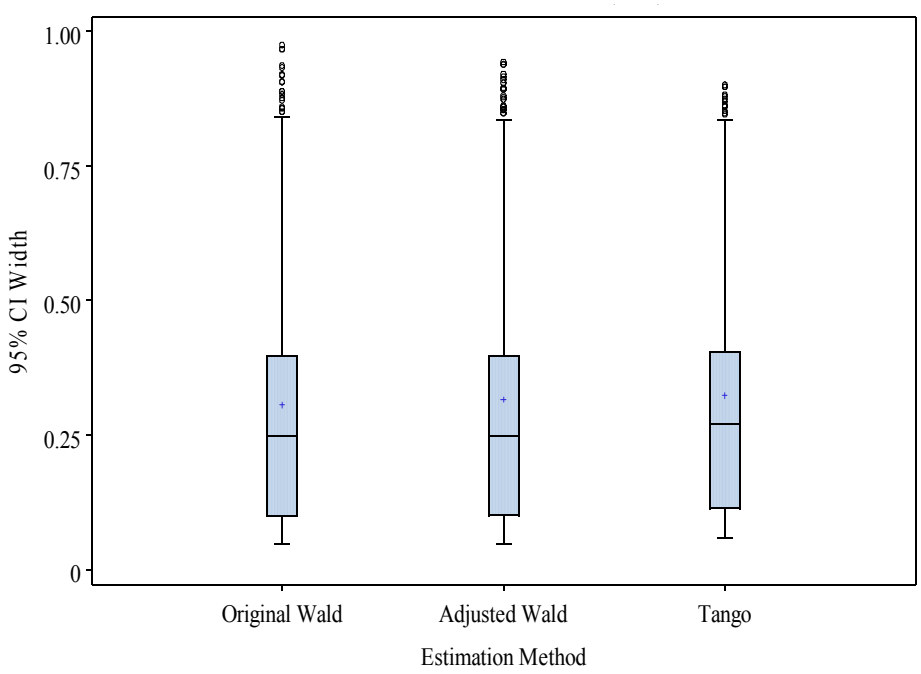

Figure 5. The distribution of interval widths at 95% confidence level.

Regarding the effect of simulation design factors on the interval width, the eta-squared associated with the main effects and the first order interaction effects were calculated at each confidence level. The results indicated that the largest effect comes from sample size with η^2 =.95. The other factors as well as interactions have relatively small η^2 values ranging from approximately 0 to .02, indicating that the variance of interval width is explained very little by the other factors in the study.

Table 8 presents the marginal means of interval width by samples size for the three methods. As indicated in the table, the interval width is very similar across the three methods at each sample size. As expected, the greatest differences in mean width are evident with the smallest sample size (and these widths with N = 10 are so large that the confidence intervals are relatively non-informative) and the interval widths decrease significantly when the sample size increases.

Table 8: Mean interval width by sample size for three methods

	Method		
Sample size	Original Wald	Adjusted Wald	Tango
10	0.74	0.79	0.77
50	0.35	0.36	0.38
100	0.25	0.25	0.27
500	0.11	0.11	0.12
1000	0.08	0.08	0.09

Note. Estimates are based on 100,000 samples of each condition.

4. Discussion

The adjusted Wald CI provided the best coverage across the conditions investigated and is easier to calculate than the Tango intervals. However, slight over coverage was observed when $N = 10$. The adjustment using $n + 2$ in the adjusted Wald CI possibly leads to over coverage when sample size is small with a greater impact of the addition of 2 to relatively small n. On the other hand, both the original Wald CI and the Tango CI produced substantial under coverage in some small sample conditions. As sample size increased, the interval coverage provided by all three methods was close to the nominal coverage level. From the perspective of Bradley's robustness criterion, however, the Tango method evidenced notably fewer conditions with adequate coverage than the adjusted Wald method. The Tango method became less adequate (likely over coverage) when the difference in two population proportions became larger and the correlation between two proportions was positive. All methods provided intervals of approximately the same average width, indicating that only minimal differences in precision are associated with the choice of method.

The results of this study clearly indicated that the original Wald method does not provide adequate coverage for smaller sample sizes. Specifically, difference between $N = 50$ and $N = 10$ were striking. Additional research should be conducted using samples sizes of 20, 30 and 40 to investigate the nuanced issues related to sample size and these methods. Even with this limitation, the results of this study suggest that the adjusted Wald CI method provides the best coverage for differences between correlated proportions.

5. References

Bonett, D. G., & Price, R. M. (2011). Adjusted Wald confidence interval for a difference of binomial proportions based on paired data. *Journal of Educational and Behavioral Statistics, 37*(4), 479-488. doi: 10.3102/1076998611411915

Bradley, J. V. (1978). Robustness? *British Journal of Mathematics and Statistical Psychology, 31*, 144-152.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157.

Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine, 17*(22), 2635-2650. doi: 10.1002/(SICI)1097-0258(19981130)17:22<2635::AID-SIM954>3.0.CO;2-C

Robey, R., & Barcikowski, R. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematics and Statistical Psychology*, 45,283-288.

Tango, T. (1999). Improved confidence intervals for the difference between binomial proportions based on paired data by robert G. newcombe, statistics in medicine, 17, 2635-2650 (1998). *Statistics in Medicine, 18*(24), 3511-3513. doi: 10.1002/(SICI)1097-0258(19991230)18:24<3511::AID-SIM303>3.0.CO;2-A

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine, 17*(8), 891-908. doi: 10.1002/(SICI)1097-0258(19980430)17:8<891::AID-SIM780>3.0.CO;2-B

Wu, P., Rodriguez de Gil, P., Pham, T., Nguyen, D., Romano, J., Kromrey, J. D. & Kim, E. S. (2013, October). *SAS® Macros CORR_P and TANGO: Interval Estimation for the Difference between Correlated Proportions in Dependent Samples*. Paper presented at the annual meeting of the Southeast SAS Users Group, St. Petersburg Beach, FL.