The Impact of Multiple Endpoint Dependency on the Q Statistic in Meta-analysis

**Purpose**

This paper examines impact of dependence on tests of homogeneity from two approaches to meta-analysis: univariate and multivariate. In univariate meta-analysis each study in a collection of studies is assumed to provide an independent effect size. In multivariate meta-analysis some or all studies provide multiple effect sizes and the correlations among the outcomes are incorporated into the meta-analysis.

We are interested in what happens when researchers violate the independence assumption by including multiple dependent effect sizes in univariate analyses. How might this practice affect meta-analytic results? The primary purpose of this paper is to analyze the effects of multiple-endpoint dependency on the homogeneity statistic $Q$. We simulate dependent data under varying conditions and analyze it in two different ways: using a typical univariate approach that ignores dependence, and using a generalized least square approach that accounts for dependence. Our rationale for examining $Q$ is threefold: it is commonly used, simple to calculate, and is often used as the primary means of choosing between fixed- or random-effects models.

**Methods/Research Design**

Consider a set of $k$ studies with one treatment group and one control group. We denote the total sample size from study $i$ as $N_i$, and the respective within-study sample sizes for treatment and control groups as $n_i^{\mathrm{T}}$ and $n_i^{\mathrm{C}}$. For simplicity we examined the case of equal sized treatment and control groups and where all individuals provide scores for $P = 2$ outcomes.

*Standardized Mean Difference*

The unbiased standardized-mean-difference (USMD) effect size (Hedges, 1981) is defined as

$$d_{ip} = \left(1 - \frac{3}{4(n_i^T + n_i^C - 2) - 1}\right)\frac{\bar{Y}_{ip}^T - \bar{Y}_{ip}^C}{S_{ip}}, \tag{1}$$

where $\bar{Y}_{ip}^T$ and $\bar{Y}_{ip}^C$ are the respective sample treatment and control means for outcome $p$ of study

and $S_{ip}$ is a sample pooled standard deviation. Hedges (1981) approximates the variance of $d_{ip}$,

$$v_{ip} = \frac{n_i^T + n_i^C}{n_i^T n_i^C} + \frac{d_{ip}^2}{2(n_i^T + n_i^C)}. \tag{2}$$

For a collection of studies, the fixed-effect mean effect size for outcome $p$ is calculated as

$$\bar{d}_{\cdot p} = \frac{\sum_{i=1}^k v_{ip}^{-1} d_{ip}}{\sum_{i=1}^k v_{ip}^{-1}}. \tag{3}$$

Last, the sample effect-size covariance for two USMD effect sizes for study $i$ is

$$cov(d_{i1}, d_{i2}) = \left(\frac{1}{n_i^T} + \frac{1}{n_i^C}\right)r_i + \left(\frac{d_{i1}d_{i2}}{2(n_i^T + n_i^C)}\right)r_i^2, \tag{4}$$

where $r_i$ is the common within-group sample correlation between outcomes from the observed

data (Gleser & Olkin, 2009).

*Heterogeneity and Univariate Q*

Statistical heterogeneity occurs when true effects differ among studies (Higgins &

Thompson, 2002). This paper looks at the effect of multiple-endpoint dependency on the widely-

used homogeneity statistic $Q$. Two forms of $Q$ were analyzed: univariate and multivariate. The

sample univariate $Q$ for outcome $p$ is defined as

$$Q_p = \sum_{i=1}^k \frac{\left(d_{ip} - \bar{d}_{\cdot p}\right)^2}{v_{ip}}, \tag{5}$$

where $k$ is the number of effect sizes. Equation (5) assesses the homogeneity of effect sizes across all studies for a single outcome $(p)$. Under the fixed-effects model and the assumption of independence of the $k$ effect sizes, $Q \sim \chi^2_{k-1}$. The statistical properties of the univariate $Q$ statistic have been well-studied under a variety of conditions (e.g., Hardy & Thompson, 1998; Harwell 1997), but all studies have assumed effect-size independence. This paper analyzes the behavior of the $Q$ statistic under a specific set of dependence conditions.

*GLS for Meta-analysis and Multivariate Q*

Consider a typical linear regression model for pairs of effect sizes of the form

$$\mathbf{d} = \mathbf{X}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\mathbf{d}$ is a $Pk$ x 1 observed effect-size matrix, $\mathbf{X}$ is a $Pk$ x $P$ matrix with elements $x_{m,n} \in \{0,1\}$, $\boldsymbol{\delta}$ is a $P$ x 1 vector whose elements are common effect-size parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a $Pk$ x 1 vector of residuals.

When the independence assumption is violated, as is the case with dependent effect sizes in meta-analysis, other estimation procedures are needed beyond the commonly-used ordinary least squares method. Generalized least squares or GLS estimation permits both heterogeneity of error variances and correlation among errors. This estimation method incorporates a variance-covariance matrix, as shown by Raudenbush, Becker, and Kalaian (1988),

$$\hat{\boldsymbol{\delta}}_{\mathrm{GLS}} = (\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^{-1}\mathbf{d}, \tag{7}$$

where

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_k \end{bmatrix} : \mathbf{S}_i = \begin{bmatrix} v_{i1} & cov(d_{i1}, d_{i2}) \\ cov(d_{i1}, d_{i2}) & v_{i2} \end{bmatrix}. \tag{8}$$

Last, we calculate the multivariate $Q$ statistic as

$$Q_M = (\mathbf{d} - \mathbf{X}\hat{\boldsymbol{\delta}}_{GLS})' \mathbf{S}^{-1} (\mathbf{d} - \mathbf{X}\hat{\boldsymbol{\delta}}_{GLS}) \tag{9}$$

The $Q_M$ statistic also follows an asymptotic chi-square distribution but with degrees of freedom equal to $df_{Q_M} = \dim(\mathbf{d}) - \dim(\hat{\boldsymbol{\delta}}_{GLS})$ (Gleser & Olkin, 2009).

Two separate simulations (*univariate* and *multivariate*) were conducted. Conditions for both simulations are shown in Table 1. For the univariate simulation we created three different sets of independent and dependent effect sizes to manipulate the level of dependence in the analysis. The multivariate simulation used GLS estimation to account for multiple-endpoint dependency. Both simulations we analyzed the impact of dependence on the homogeneity statistic by examining Type I error rates.

*Data Generation*

We generated two scores for $N_i/2$ members each in a treatment group and a control group. Scores in study $i$ were $\mathbf{Y}_i^T \sim \text{MVN}(\boldsymbol{\mu}^T, \boldsymbol{\Sigma})$ for the treatment group and $\mathbf{Y}_i^C \sim \text{MVN}(\boldsymbol{\mu}^C, \boldsymbol{\Sigma})$ for the control group, where $\boldsymbol{\mu}^T = [\mu_1^T \quad \mu_2^T]$, $\boldsymbol{\mu}^C = [\mu_1^C \quad \mu_2^C]$, and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{bmatrix}$. The population covariance and correlation of the two outcomes is $\sigma_{12}$. The standardized-mean-difference parameter for the first outcome is $\delta_1 = (\mu_1^T - \mu_1^C)/\sigma_1 \equiv \mu_1^T$, and similarly for $\delta_2$.

We implemented a correlation structure to produce multiple-endpoint type data,

$$\boldsymbol{\Sigma} = \boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{T_1 T_2} & 0 & 0 \\ \rho_{T_2 T_1} & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_{C_1 C_2} \\ 0 & 0 & \rho_{C_2 C_1} & 1 \end{bmatrix}, \tag{10}$$

where $\rho_{T_1 T_2}$ and $\rho_{C_1 C_2}$ are the population between-outcomes correlations for the treatment and control groups. For simplicity, $\rho_{T_1 T_2} = \rho_{C_1 C_2}$.

*Univariate Procedures*

Effect sizes and variances were calculated using (1) and (2), respectively. Fixed-effect means were then calculated using (3). From the information above, univariate $Q$ statistics were calculated using (5). A set of five factors were varied (see Table 1).

For the treatment group, three population outcome vectors were considered: $\boldsymbol{\delta} = \boldsymbol{\mu}^{\mathrm{T}} = [0 \quad 0] \cup [0.4 \quad 0.4] \cup [0.8 \quad 0.8]$. The population vector for the control group was always $\boldsymbol{\mu}^{\mathrm{C}} = [0 \quad 0]$, thus each pair of $\boldsymbol{\mu}^{\mathrm{T}}$ and $\boldsymbol{\mu}^{\mathrm{C}}$ provides a different magnitude of treatment effect, $\boldsymbol{\delta}$. The between-outcomes correlation $(\rho)$ measures the degree of relatedness of the outcomes. Moderate to very large correlations were considered. The number of studies $(k)$ represents small to quite large meta-analyses. Study sample size $(N)$ refers to the size of the studies within a meta-analysis, and varies from quite small to moderately large.

The dependency structure condition was the main focus of our univariate analyses (see Figures 1a-1c). These structures (*Independent*, *Moderately Dependent*, and *Very Dependent*) created sets of effects from simulated data that included dependent effect sizes. Each structure refers to groups of effect sizes with varying degrees of dependence that were univariately analyzed together.

In Figure 1a, effect sizes *within* columns come from independent studies and are statistically independent of each other. Pairs of effect sizes *across* columns come from the same studies and are correlated to a specified degree $(\rho)$. Effect sizes in univariate meta-analyses are typically analyzed separately for each outcome. When we analyze data within each column of Figure 1a, we have two independent analyses of independent effects.

Figure 1b was created by swapping the first fourth of the effect sizes from column 2 in Figure 1a with the last fourth of effect sizes from column 1, creating a somewhat elevated level

of dependence in both columns. Swapped effect sizes are outlined in bold. Half of the effect sizes in each column are dependent $\left(\{(d_{1,1}, d_{1,2}), \dots, (d_{a,1}, d_{a,2})\} \cap \right.$

$\left.\{(d_{c+1,1}, d_{c+1,2}), \dots, (d_{k,1}, d_{k,2})\}\right)$. This scenario to presents a moderate amount of dependence. Figure 1c was created using the same procedure, however the first half of the effect sizes from column 2 in Figure 1a were swapped with the last half of effect sizes from column 1. This created an even higher level of dependence in both columns compared to Figure 1b.

Frequencies of $Q$ statistics were calculated and compared to expected frequencies specified by $\chi^2_{k-1}$. With $N_{\mathrm{meta}} = 3000$ replications, the 95% confidence interval for the proportion of empirical Type I errors was $.05 \pm Z_{.05/2}\sqrt{.05(1 - .05)N_{\mathrm{meta}}^{-1}} \equiv [.042, .058]$.

*Multivariate Procedures*

The univariate approach used to analyze the data in the previous section ignored existing dependence for $d$s in two of the dependency structures. All factors except the dependency structure from univariate analyses were relevant to the multivariate analyses. A new factor (correlation type) was examined as well.

Initial steps of the multivariate procedures paralleled those described above. However, data were not reorganized as in Figures 1b and 1c because the effects for both outcomes were analyzed simultaneously using the GLS approach. An estimate of the covariance between outcomes is required. Two values were used for the Pearson product-moment correlation, $r$. First, an empirical estimate of the pooled within-groups correlation, $\bar{r}_{\mathrm{TC}}$, was calculated as

$$\bar{r}_{\mathrm{TC}} = \frac{r_{\mathrm{T}} + r_{\mathrm{C}}}{2}, \tag{11}$$

where $r_{\mathrm{T}}$ and $r_{\mathrm{C}}$ were the sample correlations between treatment and control groups. The value of $\bar{r}_{\mathrm{TC}}$ was then used to compute (4). Second, we computed subsequent GLS analyses using the

known value of $\rho$.

## Results/Findings

*Univariate Results*

Figures 2-4 present results for univariate analyses, grouped by $\delta$. Throughout all conditions under the independent structure, error rates never rose above nominal levels. When error rates were statistically different from $\alpha = .05$, they were below the nominal level. For many cases where $N = 20$, $Q$s based on independent data showed error rates that were significantly lower than expected. Harwell (1997), among others, discussed how the $Q$ statistic has low power for small studies. Error rates lower than $\alpha$ when $N = 20$ are not unexpected.

Conditions with moderately dependent data tended to produce error rates larger than those for similar conditions with independent effects. When $\rho = .50$, the majority of the error rates were within nominal levels. As the between-outcomes correlation increased, error rates increased. When the between-outcomes correlation was very strong ($\rho = .99$), nearly all error rates for conditions under the moderately dependent structure were significantly larger than expected, the largest being .088.

Trends in the error rates under the very dependent structure were similar to those of the moderately dependent structure, but with even larger empirical significance levels. The largest error rate was .107, slightly more than twice the size of the expected. This occurred for the case with the largest values of $\delta$, $\rho$, $k$, and $N$. All combinations of factors with very dependent data and strong between-outcomes correlations ($\rho = .99$) produced statistically significant error rates. When outcomes are highly correlated and analyzed together, error rates become inflated regardless of study sample size or the number of studies. Results did not appear to vary as a function of the size of $\delta$.

7

*Multivariate Results*

Figures 5-7 present error rates for multivariate analyses, grouped by $\delta$. Results for $Q_M$ differed radically from those for $Q$. Except for conditions with small studies ($N = 20$), the vast majority of error rates were within nominal levels. Overall, results for nominal error rates across most conditions were expected as GLS is a theoretically justifiable method that accounts for dependent data. As with the univariate results, results for $Q_M$ did not appear to vary by $\delta$.

Recall that for the empirical correlation conditions we calculated correlation coefficients from the observed data, while for the population conditions the known between-outcomes correlation $\rho$ values were used. For some conditions the error rates were indistinguishable by correlation type. However, in other cases error rates appeared quite different. These discrepancies were more predominant with small study sample sizes and are likely due in part to sampling error in the estimate of $\rho$. In general, empirical error rates were smaller when $Q_M$ was computed using the known value of $\rho$.

## Implications/Conclusions

The main purpose of this paper was to investigate the impact of multiple-endpoint dependency on the $Q$ statistic under a variety of conditions. The application of typical univariate analyses to dependent effect sizes led to highly inflated Type I error rates for the $Q$ statistic. Thus, if dependent data are synthesized together, the univariate $Q$ test is conservative, rejecting with rates up to twice the nominal level. This result held regardless of individual study sample size or number of studies. Results from the multivariate simulations provided evidence that the GLS method of analyzing multivariate effects successfully accounts for multiple-endpoint dependency. Nominal levels of Type I error rates were found for all conditions except for those with small studies. When empirical $\alpha$ levels deviated from the nominal .05 rate they were almost

always too low. This means the $Q_M$ test is lenient, in that it is finding heterogeneity at less than chance levels.

This work adds to the current discussion of dependence in meta-analysis. The univariate results showed how homogeneity measures are affected when multiple-endpoint data are treated as independent in meta-analysis. Our multivariate results, in contrast, demonstrated that the GLS method for meta-analysis successfully accounts for multiple-endpoint dependency.

References

Gleser, J. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. M. Cooper, L.V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. (pp. 357-376). New York: Russell Sage Foundation.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*, 841-856.

Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods*. *2*(2), 219-231.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107-128.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine*, *21*, 1539-1558.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*, 111-120.

Table 1

*Simulation Factors*

| Factor | Levels |
| --- | --- |
| Overall Effect Size ($\delta$) | 0, 0.4, 0.8 |
| Between-Outcomes Correlation ($\rho$) | .50, .75, .99 |
| Number of Studies ($k$) | 12, 24, 48, 96 |
| Study Sample Size ($N$) | 20, 100, 180, 260 |
| Dependency Structure* | Independent, Moderately Dependent, Very Dependent |
| Correlation Type ($r$)** | Empirical ($\bar{r}_{TC}$), Population ($\rho$) |

*Univariate approach only.

**Multivariate approach only.

| | |
|---|---|
| $d_{1,1}$ | $d_{1,2}$ |
| ⋮ | ⋮ |
| $d_{a,1}$ | $d_{a,2}$ |
| $d_{a+1,1}$ | $d_{a+1,2}$ |
| ⋮ | ⋮ |
| $d_{b,1}$ | $d_{b,2}$ |
| $d_{b+1,1}$ | $d_{b+1,2}$ |
| ⋮ | ⋮ |
| $d_{c,1}$ | $d_{c,2}$ |
| $d_{c+1,1}$ | $d_{c+1,2}$ |
| ⋮ | ⋮ |
| $d_{k,1}$ | $d_{k,2}$ |

Figure 1a: Independent Structure[1].

| | |
|---|---|
| $d_{1,1}$ | $d_{c+1,1}$ |
| ⋮ | ⋮ |
| $d_{a,1}$ | $d_{k,1}$ |
| $d_{a+1,1}$ | $d_{a+1,2}$ |
| ⋮ | ⋮ |
| $d_{b,1}$ | $d_{b,2}$ |
| $d_{b+1,1}$ | $d_{b+1,2}$ |
| ⋮ | ⋮ |
| $d_{c,1}$ | $d_{c,2}$ |
| $d_{1,2}$ | $d_{c+1,2}$ |
| ⋮ | ⋮ |
| $d_{a,2}$ | $d_{k,2}$ |

Figure 1b: Moderately Dependent Structure.

| | |
|---|---|
| $d_{1,1}$ | $d_{b+1,1}$ |
| ⋮ | ⋮ |
| $d_{a,1}$ | $d_{c,1}$ |
| $d_{a+1,1}$ | $d_{c+1,1}$ |
| ⋮ | ⋮ |
| $d_{b,1}$ | $d_{k,1}$ |
| $d_{1,2}$ | $d_{b+1,2}$ |
| ⋮ | ⋮ |
| $d_{a,2}$ | $d_{c,2}$ |
| $d_{a+1,2}$ | $d_{c+1,2}$ |
| ⋮ | ⋮ |
| $d_{b,2}$ | $d_{k,2}$ |

Figure 1c: Very Dependent Structure.

---

[1] Values of the subscripts $a$, $b$, and $c$ were $a = k/4$, $b = k/2$, $c = 3k/4$.
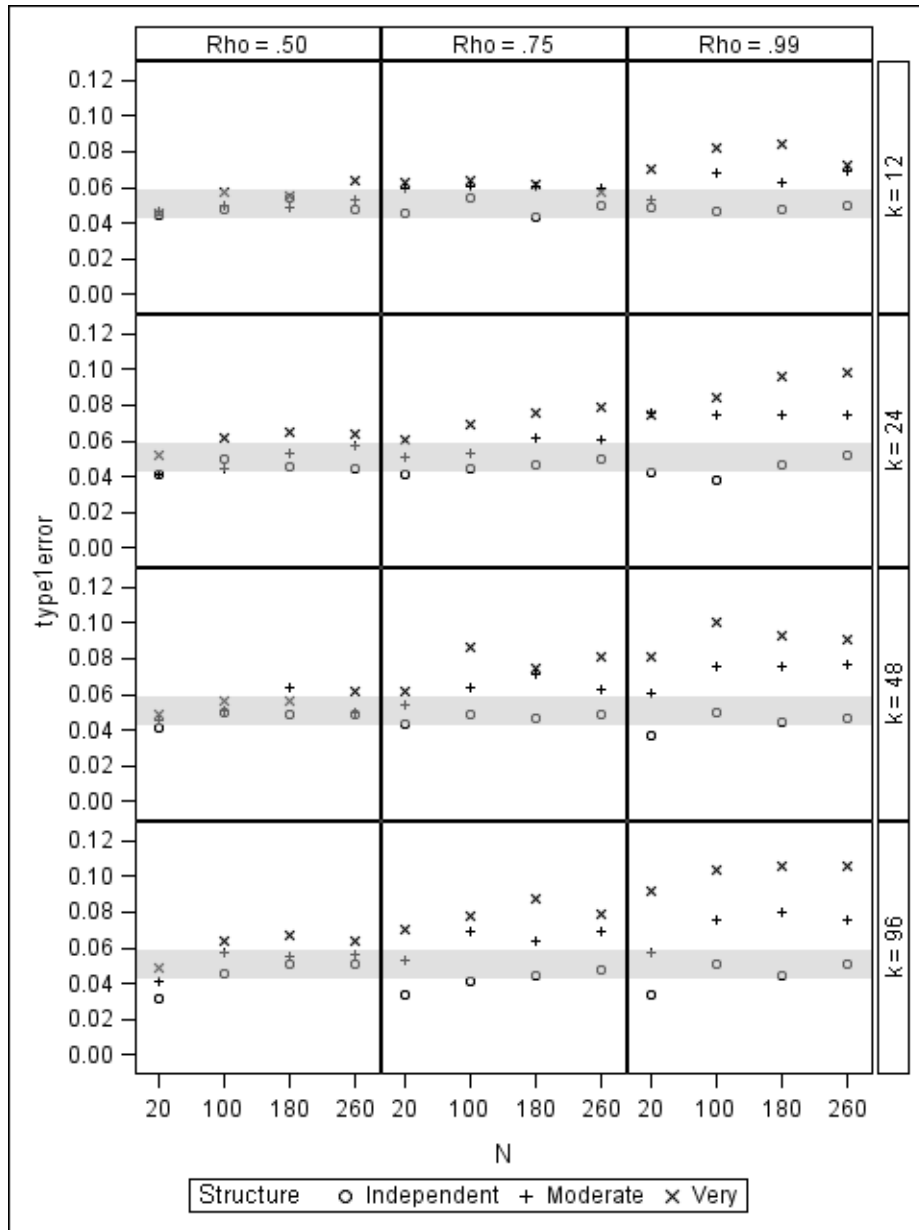
Figure 2: Univariate Results ($\delta = 0$).
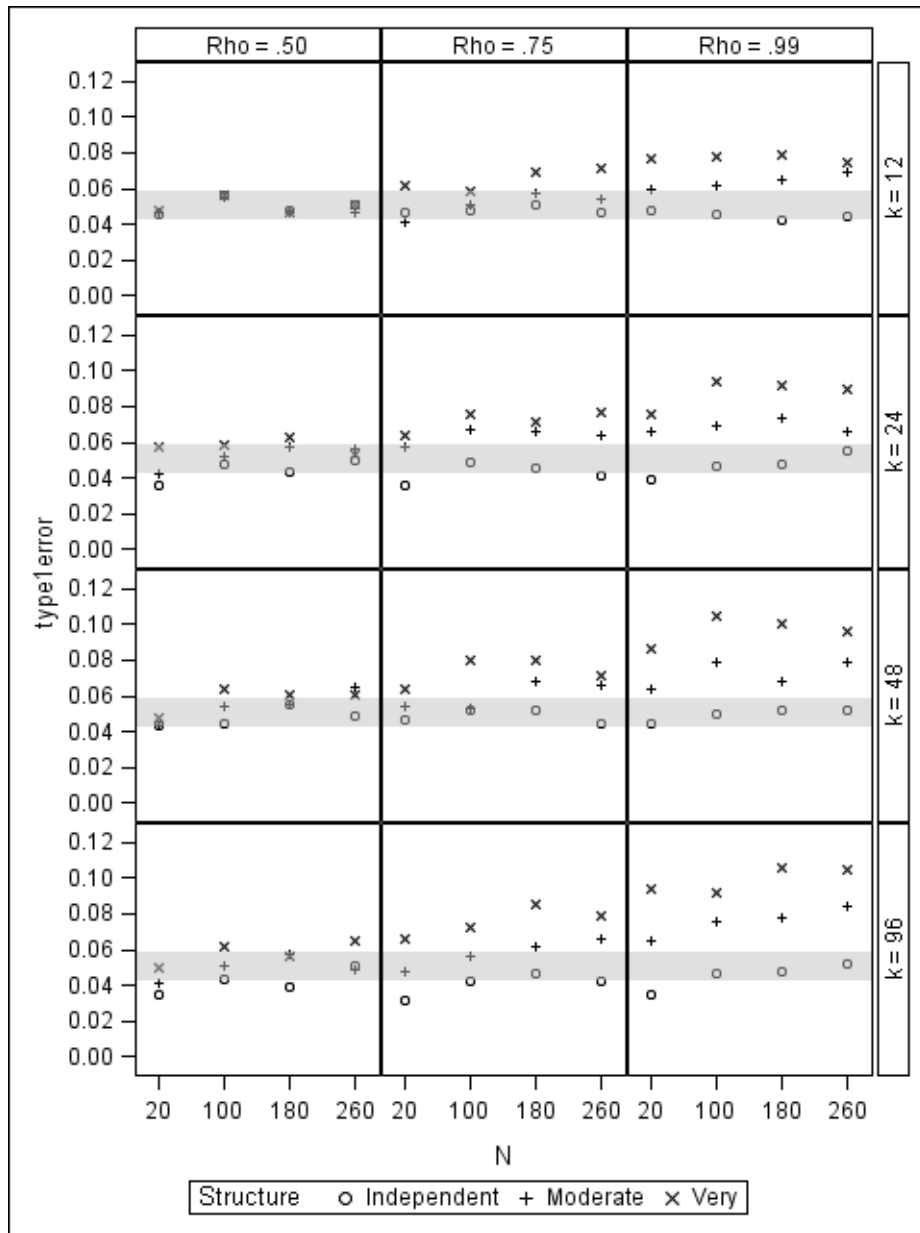
Figure 3: Univariate Results ($\delta = 0.4$).
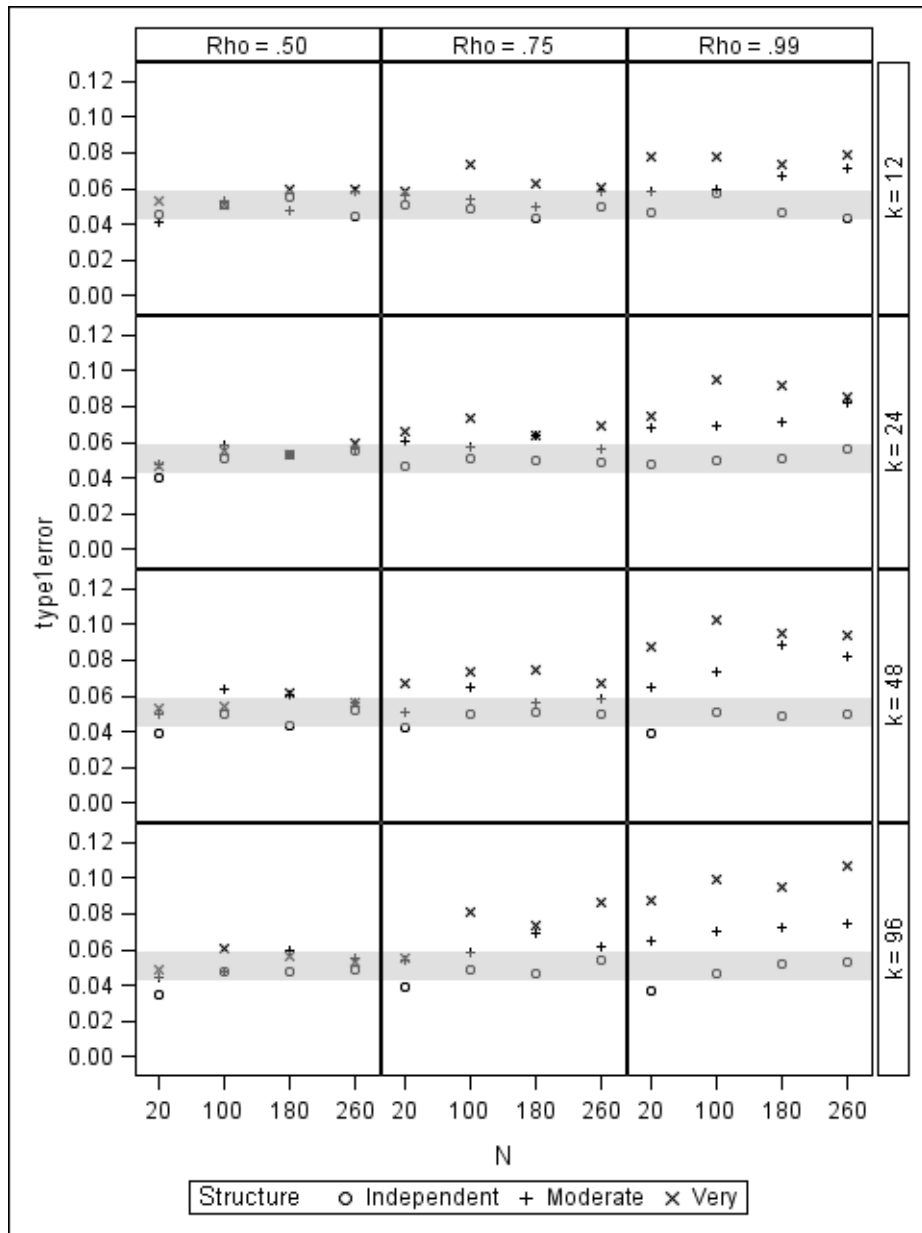
Figure 4: Univariate Results ($\delta = 0.8$).
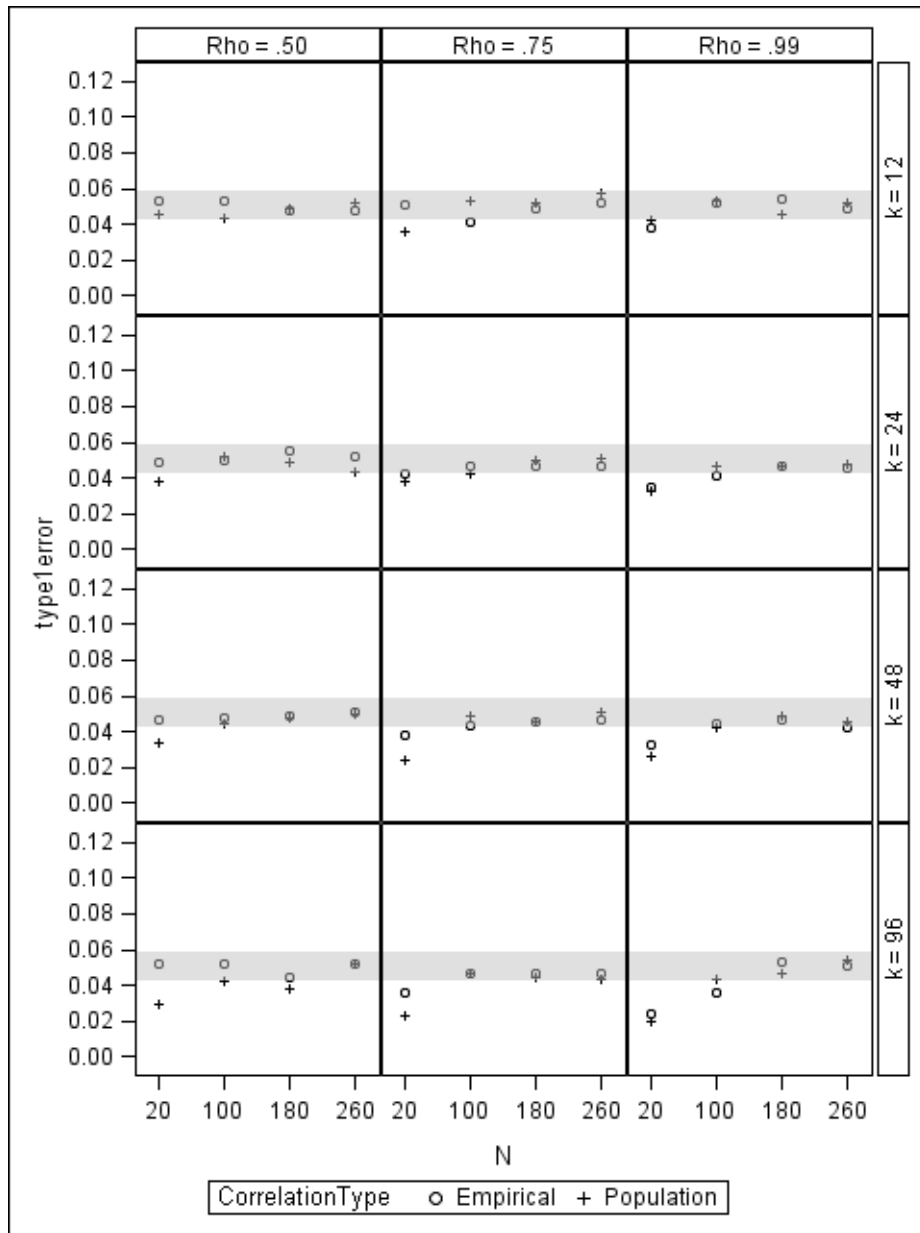
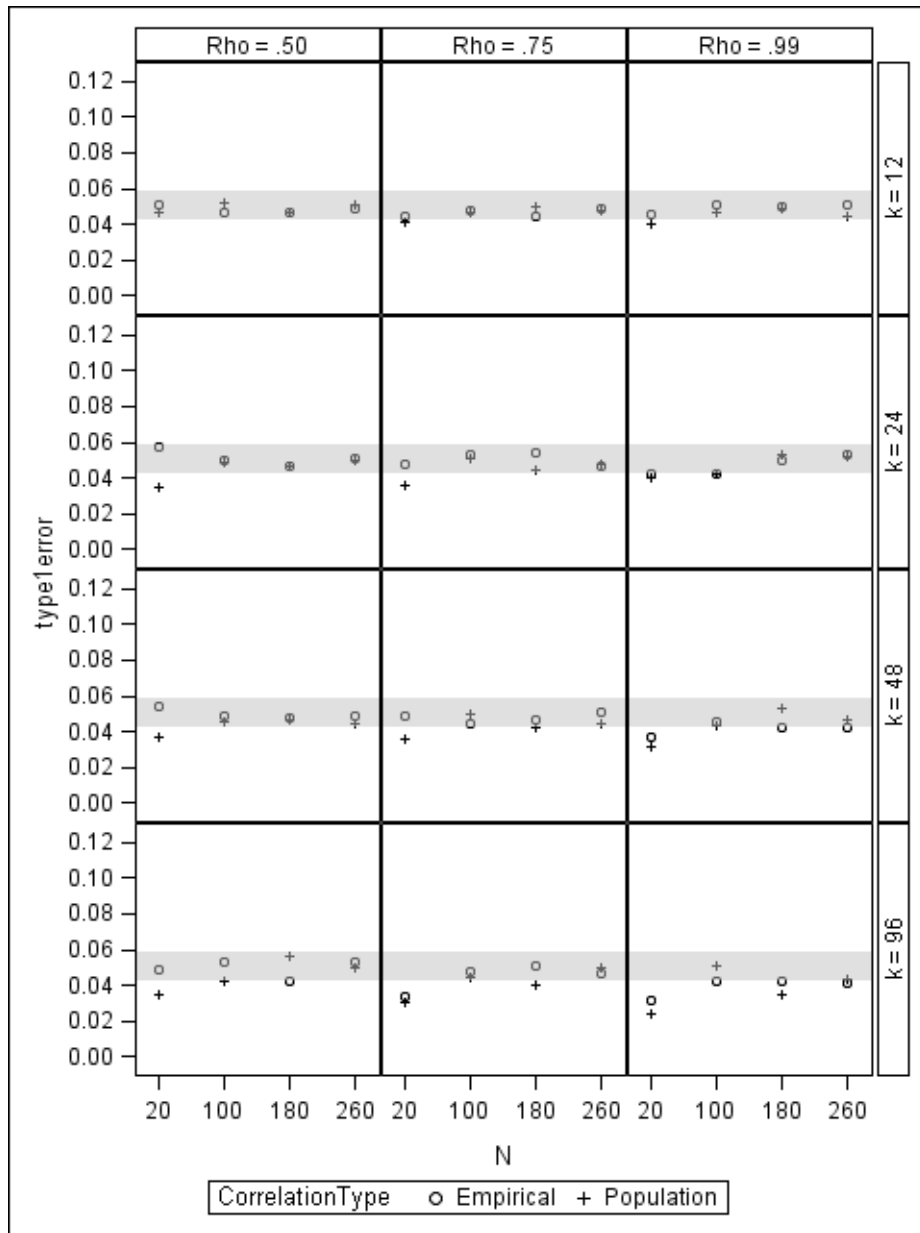Figure 5: Multivariate Results $(\delta = 0)$.
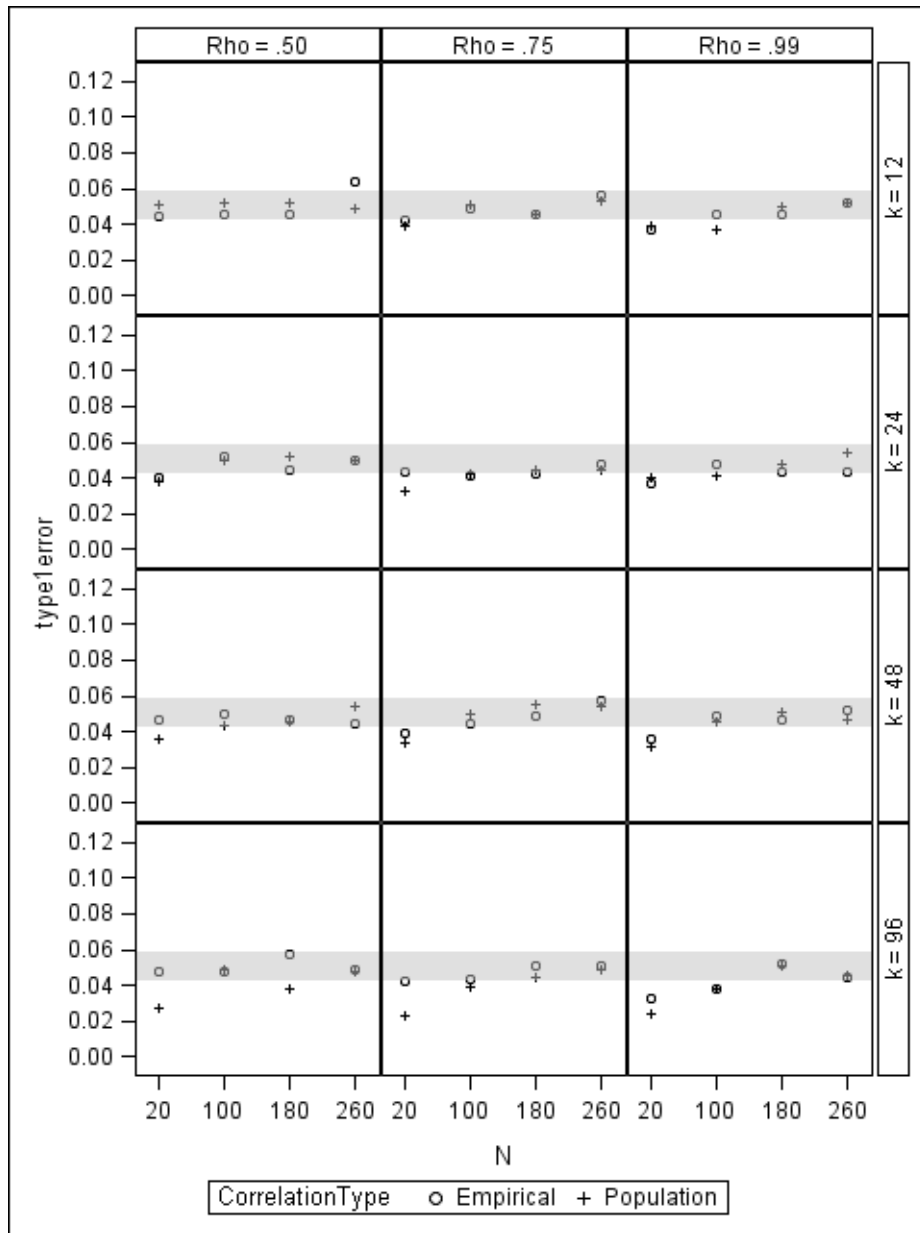
Figure 6: Multivariate Results ($\delta = 0.4$).

Figure 7: Multivariate Results ($\delta = 0.8$).