Journal of Behavioral Data Science, 2026, *6* (1), 1-40.
DOI: https://doi.org/10.35566/jbds/choiwu

Zero-Frequency Cell Correction Strategies in Tetrachoric Correlation Estimation: Expanded Strategies and Multivariate Implications

Jeongwon Choi[1 [0000−0001−6087−2124]] and Hao Wu[1 [0000−0001−6471−1774]]

Vanderbilt University, Nashville, TN 37203, USA
jeongwon.choi@vanderbilt.edu, hao.wu.1@vanderbilt.edu

Abstract. Zero-frequency cells pose a challenge for tetrachoric correlation estimation, but investigation of correction strategies remains limited. This study evaluates several zero-cell correction strategies, including different values to add, different ways to add the value, and the use of unadjusted versus adjusted thresholds in the second stage in the two-stage procedure. These strategies are examined across different correlation sizes and thresholds, to estimate a single tetrachoric correlation and extended to multivariate applications involving a tetrachoric correlation matrix and a confirmatory factor analysis model for binary data. Using multiple evaluation criteria, we show how these strategies perform differently across correlation sizes and the pattern of thresholds. This study also introduces ways to improve computational efficiency for tetrachoric correlation simulation studies that leverage the discrete structure to reduce redundant computations.

Keywords: Tetrachoric Correlation · Zero-frequency Cells · Binary Data · Tetrachoric Correlation Matrix · Confirmatory Factor Analysis

1 Introduction

1.1 Polychoric and Tetrachoric Correlations

Psychological research often produces ordered categorical data. In this situation, such variables are assumed to have arisen from discretizing multivariate-normally distributed underlying responses by thresholds. The correlations in this multivariate normal distribution are polychoric correlations. Their estimation has wide applications in structural equation modeling and item factor analysis. A special case of the polychoric correlation is the tetrachoric correlation between two binary variables.

The two-stage procedure (Olsson, 1979) is the most widely used approach for estimating polychoric correlations. It first estimates the thresholds for each variable using its marginal category proportions and then for each pair of variables

maximizes the likelihood of the proportions in their two-way contingency table as a function of the correlation, treating the thresholds as fixed. For the special case of estimating a single tetrachoric correlation, the two-stage procedure is equivalent to maximum likelihood where the two thresholds and the single correlation are estimated jointly in a single stage to maximize the likelihood of the observed proportions, because both procedures would produce estimates that can perfectly reproduce the observed proportions.

1.2 The Zero-Frequency Issue in Tetrachoric Correlation Estimation

One major issue in estimating a polychoric correlation is the presence of zero-frequency cells in the contingency table. Zero-frequency cells are common when the sample size is small (e.g., less than 200), when the estimated thresholds are extreme, or when the underlying correlation of the variable is high (Savalei, 2011).

The zero-frequency issue in 2×2 tables is distinctive because a bivariate normal distribution with a perfect correlation can exactly reproduce the observed proportions in a 2×2 table that contains one zero cell. This means that even a single zero cell pushes the best-fitting correlation to the boundary (± 1) , whereas larger tables do not have this property.

Figure 1 illustrates this point. The left panel shows the space of two latent responses that produce a 2×2 table with one zero. In this case, a degenerate bivariate normal distribution whose support is on the slanted gray line can perfectly reproduce the observed proportions once discretized by thresholds (represented by the dotted lines) calculated through the observed marginal proportions. This means the correlation of 1, being able to perfectly reproduce the observed proportion, must be the maximizer of the likelihood function, yielding a correlation estimate on the boundary of the parameter space. In contrast, for tables larger than 2×2, a single zero-frequency does not yield a boundary solution, because a bivariate normal distribution with a perfect correlation necessarily results in at least two zero cells. As shown in the right panel of Figure 1, the linear subspace (represented by the gray line) on which a degenerate bivariate normal distribution is supported can go through at most four of the six regions defined by the thresholds, leaving at least two zero-probability cells in the 2×3 table. Because a zero probability cell cannot produce a nonzero count, the analysis above means a boundary correlation of ± 1 would produce a likelihood function value of 0 for bivariate data with only one zero in a contingency table bigger than 2×2 and therefore cannot be an estimate. This difference highlights why the zero-frequency problem is uniquely severe in 2×2 tables. Motivated by this distinctiveness, in this paper we focus on the estimation of tetrachoric correlations among binary variables. Prior research (Savalei, 2011) also found that zero-frequency cell treatment is most relevant for binary data.

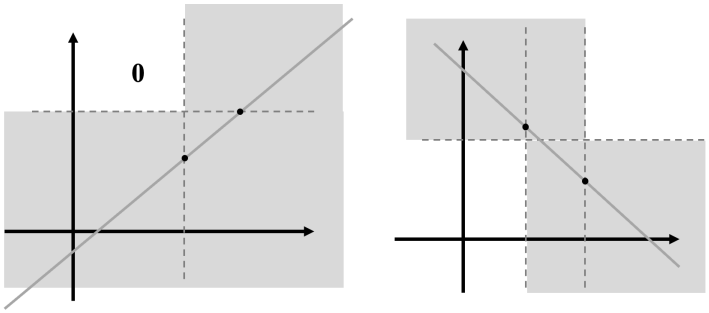


Figure 1: A degenerate bivariate normal distribution. Left: 2×2 table with one zero cell. Right: 2×3 table with two possible zero-cell locations under perfect correlation. The solid line represents a bivariate distribution with perfect correlation; dotted lines indicate thresholds.

1.3 Correction Strategies for Zero-Frequency Cells

The most widely used approach to address the problem of boundary solutions caused by zero-frequency is to add a small value to the zero-frequency cell in the contingency table. This idea was initially introduced by Brown and Benedetti (1977) for tetrachoric correlation based on the idea of Yates (1934)'s correction for continuity. This modification is available in several R (R Core Team, 2022) packages, such as the `PolychoricRM` function in the `Turbofuns` package (Zhang, Trichtinger, Lee, & Jiang, 2022), `polychoric.matrix` function in the `EGAnet` package (Golino & Christensen, 2025), the `lavCor` function in the `lavaan` package (Rosseel, 2012), and the `polychoric` function in the `psych` package (Revelle, 2023). These software programs make it easy for researchers to implement the correction method of adding a small value.

Although the general idea across software packages is to add a small constant, the specific way this correction is applied differs. First, different values can be added. While 0.5 is the most frequently used as a value to add to zero-frequency cells, other values such as 0.1 or the inverse of the total number of cells in the bivariate contingency table (0.25 for calculating a tetrachoric correlation) are also being used. For example, `Turbofuns` offers the option to use values of 0.1, 0.5, and the inverse of the number of cells. It remains unclear which value is optimal. Secondly, one can add the small number only to the zero-frequency (e.g., `lavaan`, `psych`) or to all cells (e.g., `Turbofuns`, `EGAnet`). One can even choose to add to all cells even when zero-frequency is not present within the contingency table. There is also the choice to maintain the marginal proportions (e.g., `lavaan`), which aligns with Brown and Benedetti (1977)'s idea that the remaining cells should also be adjusted according to the substitution of zero cell to maintain the marginal. However, this option is limited to 2×2 tables and is not available for larger contingency tables, which makes it especially relevant to examine in the 2×2 case. Lastly, additional inconsistencies arise in the use of

thresholds in the second stage of the two-stage procedure when zero-frequency cells are present in the contingency table. Mplus (Muthén & Muthén, 2017) and `lavaan` employ thresholds calculated from the uncorrected table (referred to as unadjusted thresholds), whereas several R programs, including `psych` and `Turbofuns`, use thresholds calculated from the table after the zero-frequency cell correction (referred to as adjusted thresholds).

1.4 Prior Research on Zero-Frequency Cell Corrections

While the zero-frequency cell issue is critical in tetrachoric correlation estimation, there has been relatively limited research on these correction strategies. Initial work on this topic carried out by Savalei (2011) focused primarily on comparing the impact of adding 0.5 with that of making no modification for zero-frequency cells. For binary data, their simulation conditions included correlations of 0, 0.3, 0.5, 0.7 and 0.9 combined with various threshold values. Their research concluded with a tentative recommendation of adding 0.5 to the zero cell over making no modification. In particular, their results showed that adding 0.5 to a zero cell yielded more converged replications, a bell shaped sampling distribution for the estimated correlation, smaller empirical standard error, less biased estimated standard error, and, for the more typical situations of small correlations, a better coverage of confidence intervals, while it was also noted that with opposite-signed extreme thresholds, neither adding 0.5 nor not adding performed well.

Yang and Weng (2024) extended Savalei (2011) by incorporating smaller thresholds when comparing two approaches: adding 0.5 only to the zero cell versus leaving zero cells uncorrected. They considered the threshold sets used in Savalei (2011) and added smaller values to represent milder skewness of the observed data. For the larger-threshold conditions (the same as in Savalei, 2011), they found that zero-cell correction was most beneficial, especially under strong correlations with thresholds in the same direction. This result was consistent with Savalei's earlier findings. In these scenarios, adding 0.5 generally produced smaller empirical standard errors, particularly when correlations were not too high (< 0.7). They also conducted a supplementary simulation study evaluating different methods for adding a number, except for smaller thresholds conditions. They reported that different correction techniques, such as adding 0.5 to all cells, adding 0.5 and keeping the marginal counts, and adding the reciprocal of the number of cells, yielded results similar to adding 0.5 only to zero cells. Meanwhile, adding the reciprocal of the number of cells to all cells resulted in lower bias but higher empirical standard error compared to adding 0.5 only to zero cells.

For smaller threshold conditions (i.e., smaller than those in Savalei, 2011), Yang and Weng (2024) recommended leaving zero cells uncorrected. However, this conclusion may reflect a natural limitation of their simulation design. As shown in their Table 3, the number of replications with zero-frequency cells is extremely small under milder threshold conditions. For example, with small and same-direction thresholds, the average number of zero-frequency cells across

datasets was zero. With opposite-direction thresholds, replications with zero-frequency cells also did not occur unless correlations were very high (0.7 or 0.9).

1.5 The Purpose and Overview of the Present Study

Despite these existing works, significant gaps remain, and this study aims to address them. First, there are additional strategy dimensions that prior work did not explore. For example, it did not consider adding 0.1 or adding a small value to all cells regardless of the presence of a zero cell. Nor did it compare the use of adjusted and unadjusted thresholds in the second stage of estimation after the adjusting the table. These strategies will be examined in this study.

Second, the behavior of standard errors has not yet been studied across different choices of the added constant for zero cells. Yang and Weng (2024) examined several zero-cell corrections but reported only empirical standard errors, not estimated standard errors. Therefore, we extend this work by assessing both point estimates and the properties of Wald confidence intervals across correction strategies.

Third, this study includes a sample size of 50 to fill the gap in research on smaller sample sizes, specifically those less than 100. It is particularly important because smaller sample sizes are more prone to the occurrence of zero-frequency cells, which can intensify the issues associated with these occurrences.

Finally, we extend the evaluation of these strategies from estimating a single correlation to settings with multiple variables, which are more relevant to practical applications. In practice, tetrachoric correlations are usually computed as a preliminary step to construct a full correlation matrix for further statistical analyses, such as structural equation modeling or item factor analysis. When zero-frequency cells are present and tetrachoric correlations are estimated pairwise, the resulting matrix can easily have non-positive eigenvalues (Deng, Yang, & Marcoulides, 2018; Yuan, Wu, & Bentler, 2011), which leads to difficulties for model fitting. For this reason, we evaluate correction strategies in multivariate settings, examining their impact on the entire correlation matrix, and, in a confirmatory factor analysis (CFA) application, on how such corrections affect model estimation results.

In addition to expanding the scope of the existing studies on zero-frequency cell correction, we also propose ways to improve the computational efficiency and accuracy of simulation studies for discrete problems. First, we reduce redundant estimation by exploiting the discrete nature of the problem and the symmetry of the model. Traditional simulations randomly generate a large number of datasets from a distribution and run competing statistical procedures on each of them to produce outcomes. For a discrete problem, this tends to generate the same dataset (i.e., 2×2 table) multiple times, leading to repeated calculations and computational inefficiency. Our approach exploits this discrete nature and certain symmetry of the problem and avoids unnecessary estimations, saving computational time. Second, we reduce simulation error by obtaining the theoretical sampling distribution instead of relying on a limited number of random replications (e.g., $1,000$ or $10,000$) under each condition. Further details

are provided in the Methods section of Study 1 under "Strategies for Efficient Simulation."

This paper presents three studies. The first study is a simulation comparing strategies for estimating a single correlation by varying both the value added and specific ways to add this number. The second extends the comparison to multivariate settings, and the third applies these strategies in a confirmatory factor analysis for binary data. We then synthesize the findings across the three studies and discuss their implications in the conclusion and discussion section.

2 Study 1: Bivariate Analysis

The first simulation study was conducted to evaluate various methods with different added values and different manners to add values for zero-cell correction to obtain a single correlation estimate.

2.1 Methods

Simulation Conditions In our simulation, we considered three factors: sample size (50, 100, 200), the magnitude of the underlying correlation (0.3, 0.5, 0.7, 0.9), and thresholds for each variable. The sets of thresholds were created with -1.5, -1.0, -0.8, 0.8, 1.0, and 1.5 based on Savalei (2011). We did not include the additional smaller thresholds considered by Yang and Weng (2024), since milder thresholds rarely generate zero-frequency cells. Instead, we focused on the threshold sets from Savalei to better represent scenarios where the zero-frequency problem is more likely to occur. Thus, the correlation and threshold values were consistent with Savalei (2011). We also introduced a sample size of 50 to reflect a situation where more zero cells are likely to exist. The lack of this sample size was mentioned as a limitation of Yang and Weng (2024), highlighting the need to analyze a sample size of 50 because empirical studies often have a smaller sample size.

We considered different zero-cell correction options used in practice and software. In addition to the default of not making correction, 12 different correction methods for zero-frequency cells were considered, involving two dimensions of corrections: which value to add (0.1, 0.25, 0.5) and how a value is added (keeping the marginal counts, only adding to the zero, adding to every cell when a zero cell is present, or always adding to all cells).

In the second stage of the two-stage procedure, the options of using unadjusted and adjusted thresholds were both considered whenever the zero-cell correction changes the marginal proportions. Adjusted thresholds are the thresholds calculated using the table corrected for zero-frequency cells, whereas unadjusted thresholds are the thresholds derived from the raw table without such correction.

Strategies for Efficient Simulation The purpose of a simulation study is typically to obtain the sampling distribution of an outcome of a statistical procedure. In this study, we minimized simulation error arising from the randomness

of the Monte Carlo procedure by calculating the sampling distribution directly. Note that due to the discrete nature of the problem, there are a large but limited number of different 2×2 contingency tables for each given sample size. There are 23,426 tables for the sample size of 50, there are 176,851 tables for the sample size of 100, and there are 1,373,701 tables for the sample size of 200.

Given each population correlation and threshold of a simulation condition, we calculated the theoretical probability for each contingency table to be sampled. Specifically, the theoretical probabilities for each of the four cells in the 2×2 table can first be computed based on discretizing the bivariate normal distribution, and then the probability of each contingency table can be computed from a multinomial distribution with the given cell probabilities and observed counts.

In theory, these probabilities of all 2×2 contingency tables define the theoretical sampling distribution of the observed contingency tables. Once tetrachoric correlation is estimated from each contingency table, these probabilities also define the theoretical sampling distribution of the tetrachoric correlations. For a sample size of 50, all possible tables except for those with two or more zero cells[1] were estimated to construct the sampling distribution. This involved a total of 23,128 tables. [2]

However, for larger sample sizes of 100 and 200, only tables with a probability of at least 10^{-5} in at least one condition were considered to exclude rare tables, reducing the total number of tables estimated. This resulted in a total of 15,829 tables (15,732 tables when those with two or more zeros were further excluded) for a sample size of 100 and 58,397 tables (58,394 tables when those with two or more zero cells were further excluded) for a sample size of 200.

The number of estimated tables can further be reduced by identifying prototype tables. This strategy was applied to sample sizes of 100 and 200. For two binary variables, because switching the two categories of either variable or switching the two variables results in predictable changes (e.g., a flip in sign) in the thresholds or correlation, there is no need to analyze every distinct 2×2 table. Rather, only one "prototype" table needs to be analyzed. Consider bi-

[1] We excluded tables with two zero-frequency cells, whether on the diagonal or on the same row or column, because either the two variables are perfectly related to each other and only one of them is retained in practice, or one variable is degenerated and needs to be removed in practice. The removed tables account for a small fraction in terms of both count and probability.

[2] For a fixed sample size N, a 2×2 contingency table can be represented by nonnegative integer cell counts (a, b, c, d) satisfying $a + b + c + d = N$. The number of such tables is $\binom{N+3}{3}$ (equivalently, ${}_4H_N$), which is the same as the number of ways to choose three balls from a sequence of 53 balls to determine the number of balls between and beyond the three chosen balls as the desired partition of 50. For $N = 50$, this gives ${}_4H_{50} = \binom{53}{3} = 23{,}426$ possible tables. We then excluded tables with two or more zero cells. The number of tables with exactly two zero cells is $\binom{4}{2}(N-1)$: there are $\binom{4}{2} = 6$ ways to choose which two cells are zero, and the remaining two positive counts must sum to N, which yields $N - 1$ possibilities. The number of tables with exactly three zero cells is 4 (all observations fall in a single cell). Thus, the number excluded is $6(N - 1) + 4$, which equals 298 when $N = 50$, leaving $23{,}426 - 298 = 23{,}128$ tables.

variate contingency tables with counts n_{00}, n_{01}, n_{10}, and n_{11}, where the two subscripts indicate the category labels for the first and the second variables. Every table can be turned into a prototype table that satisfies $n_{00} \geq n_{01} \geq n_{10}$ and $n_{00} \geq n_{11}$ by switching the categories or the two variables. Once the prototype table is analyzed, the parameter estimates can be modified (e.g., through sign changes or flips of the thresholds) to obtain the estimates for the other seven related tables.

We also simplified the number of conditions by reducing the number of correlations and thresholds considered in the study. Without loss of generality, we only included positive correlations, because changing the order of responses can account for negative correlations. With this approach, half of the possible correlations can be removed from the analysis. For thresholds, we only considered the situation where the first threshold is positive and no less than the absolute value of the second threshold. Consequently, with six different threshold values, the total number of threshold pairs was reduced to 12 from 36. Specifically, there are $6 \times 6 = 36$ ordered pairs. By imposing an ordering (threshold1 $\geq$ threshold2; with threshold1 > 0 to avoid symmetric duplicates), (threshold1, threshold2) and (threshold2, threshold1) are treated as the same, leaving $6 \times 5/2 = 15$ unique pairs. Excluding the three same-threshold cases leaves 12. When fully crossing these three factors, a total of 144 conditions (3 sample sizes $\times$ 4 correlations $\times$ 12 thresholds) were obtained. This is significantly smaller than the number of conditions without reduction, which would have been 864 (3 sample sizes $\times$ 8 correlations $\times$ 36 thresholds).

Computation After generating a table, we computed the correlation estimate ($\hat{\rho}$) and its standard error for each prototype table through the two-stage procedure using our modified version of the function `polychor` in the `polycor` package (Fox, 2022). Moreover, when a non-excluded table contained a zero cell, we did not estimate the uncorrected correlation; instead, we set it to 1 or -1 (depending on the location of the zero cell), because these are the theoretical boundary values that maximize the likelihood function at the second stage (see Figure 1). The use of iterative estimation for such tables typically results in a value close to 1 or -1 that depends on the optimization package used. These approaches were also used in the later simulations in Study 2 and Study 3.

The loss function minimized in the second stage is defined as L below. In the formula, N is the total sample size, n_{ij} is the count in nonzero cells where i and j correspond to response categories for the first and second variables, respectively, and p_{ij} is the expected proportion of each cell:

$$L = -\frac{1}{N} \sum_{j=0}^{1} \sum_{i=0}^{1} n_{ij} \cdot \log \left(\frac{p_{ij}}{n_{ij}/N} \right). \tag{1}$$

We used the Nelder–Mead algorithm in the `optim` function in R to estimate $\hat{\rho}$ by minimizing the loss function L, which is defined on $[-1, 1]$ but set as undefined beyond -1 and 1. The R code used in our computations is available in the OSF repository: `https://osf.io/hmk2e/`.

Evaluation Criteria We used three different criteria to evaluate point estimates of a single tetrachoric correlation: the root mean square error (RMSE) and the mean absolute error (MAE) for correlation estimates, and the MAE of Fisher's z-transformed correlation estimates. These measures were calculated using the probabilities of the tables being sampled as weights.

Fisher's z-transformation is particularly relevant for the boundary-solution problem, because it amplifies the penalty for near-perfect correlations. This makes Fisher's z an especially important complement to RMSE and MAE, which evaluate errors uniformly across the correlation range without giving extra weight to boundary cases. Fisher's z-transformation for the correlation is defined as follows.

$$z = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) \tag{2}$$

After applying this transformation, the transformed values were used in place of the raw correlation coefficient $(\hat{\rho}, \rho)$ when computing error. Fisher's z-transformation cannot be used with a perfect correlation, so it was not applied when no modification was used for zero counts.

The estimated standard error (SE) is typically used to form a Wald confidence interval (CI) as an interval estimate, so to evaluate the SE we computed the noncoverage rates of the 95% Wald CI of the correlation. Because when a zero-frequency cell is present but no correction is made, the estimate must be 1 or -1 and the SE cannot be properly estimated, the noncoverage rates were only calculated when a nonzero added value was used.[3]

2.2 Results

In this section we present results for the sample size of 50. The sample sizes of 100 and 200 led to similar patterns in the results, and their difference from the sample size of 50, if present, will be noted below in footnotes. The tables for all sample sizes and figures for the two larger sample sizes are provided in the Appendix. In the figures, rows are ordered by increasing correlation size from top to bottom, while columns are ordered by the distance between thresholds, with thresholds becoming closer from left to right. In the far-right three columns with identical thresholds, the order is based on the size of the threshold, with the more extreme threshold on the left.

The Number of Zero-Frequency Cells Table A1 in the Appendix presents the composition of the sampling distribution of 2×2 tables based on the number of zero-frequency cells. This includes scenarios with no zero-frequency cells, one zero-frequency cell, and two or three zero-frequency cells. The sampling distribution primarily consists of tables with no zero-frequency cell or just one, while

[3] When the population correlation is 1 (or -1), there must be at least one zero cell in the contingency table and the estimate must be 1 (or -1), so the theoretical SE is zero.

probabilities of two or more zero-frequency cells are relatively rare, and such tables were removed from the analysis. Specifically, in 52.08% of conditions (25 out of 48), the probability of the presence of exactly one zero is greater than 0.5, while only in 8.33% of conditions (4 out of 48), this probability is less than 0.05.[4] Zero-frequency cells appear more often in cases with high correlations or extreme thresholds. For example, high correlations like 0.7 or 0.9 combined with thresholds of opposite signs lead to higher probabilities of one zero cell. These results indicate our choice of simulation conditions has properly captured the scenarios with zero-frequencies.

Root Mean Square Error (RMSE) The RMSE values of the correlation estimates were calculated for each combination of threshold sets, correlation sizes, and different modifications for zero-frequencies. Modifications included which value to add, how to add it, and whether to use adjusted or unadjusted thresholds in the second stage of the two-stage procedure. Figure 2 presents the RMSE for point estimates of the correlation when the sample size is 50.

The overall pattern in Figure 2 indicates that the added value plays the primary role in determining the results, and that, given the optimal added value, the manner of addition has only a minor influence. The choice of unadjusted or adjusted thresholds makes little difference. For the added value, there is a general trend that as thresholds become far apart and correlation becomes greater, adding a smaller number tends to produce the best result. In particular, when thresholds have the same sign (i.e., the right block of Figure 2), the use of 0.5 as the added value appears to be optimal in all panels, with possible exceptions for the highest correlation and most distant thresholds, for which smaller added values may be optimal. For opposite-signed thresholds and the highest correlation (i.e., last row of the left block), not adding a number to the zero cells produces the lowest RMSE. In the remaining panels of this figure (i.e., first three rows of the left block), the optimal added value increases from 0 to 0.5 from the lower left to the upper right.

The mean absolute errors (MAE) show a similar pattern to RMSE. The relevant figures can be found in the Appendix (Figures A1, A4, A5).

Mean Absolute Error of Fisher's Z-Transformed Correlation Figure 3 presents the MAE of Fisher's z-transformed correlation estimates for a sample size of 50. It shows that the MAE for Fisher's z-transformed correlation estimates exhibits a very similar pattern, but no modification is no longer an option because it would produce an estimate of 1 or -1 and an infinite transformed value. Although the added value drives the general pattern, under some conditions the

[4] For a sample size of 100, 45.8% (22 out of 48 conditions) have a probability of exactly one zero greater than 0.5, while in 22.9% (11 out of 48) this probability falls below 0.05. For a sample size of 200, in 37.5% (18 out of 48) conditions this probability exceeds 0.5, and in 47.9% (23 out of 48) conditions it falls below 0.05 (see Tables A2 and A3 in the Appendix).

specific way of adding the value leads to notable differences within the same added value. For instance, for correlations of 0.3 and 0.5, and thresholds of 1.5 and 1, keeping the marginals and adding to all cells regardless result in much lower MAE within an added value of 0.5 compared to other methods. When thresholds have opposite signs, the optimal added value decreases from 0.5 to 0.1 as the thresholds become farther apart and the correlation becomes greater. With the optimal added value, the way to add does not matter much. When thresholds are of the same sign, in most cases (with the possible exception of a correlation of 0.9 combined with thresholds 1.5 and 0.8), adding 0.5 produces the smallest MAE. With this optimal added value, either adding it consistently to all cells even when no zero is present or adding it while keeping the marginals is among the best ways to add.

Noncoverage Rates The noncoverage rates for the 95% Wald CI are provided in Figure 4. We compared different correction methods against the nominal level of 0.05. When there are zero cells in the table, standard errors cannot be computed without zero-cell correction due to the boundary estimate of ± 1. Therefore, coverage rates were not calculated for no correction. Because adjusted and unadjusted thresholds produced very similar point estimates, only adjusted thresholds were considered in the calculation of SE and CI.

In scenarios with positive thresholds, adding 0.5 leads to the lowest noncoverage rates in almost all conditions. Specifically, when thresholds are not extreme, adding 0.5 to all cells regardless of the presence of zero cells tends to result in the lowest noncoverage rates, which are also the rates closest to the nominal level of 0.05. In cases with at least one extreme threshold of 1.5, adding to zero cells while maintaining the marginals tends to give the lowest noncoverage rates, which are also the rates closest to the nominal level for small and moderate correlations; for higher correlations, a smaller added value with marginals maintained may lead to the closest noncoverage to the nominal value of 0.05.

When thresholds have mixed signs, adding different values generally produces very similar results, which are mostly below 0.05, except for some combinations of high correlation or extreme thresholds. For small correlations and closer thresholds, adding a larger value (0.5) is a better strategy compared to adding smaller values. However, when a high correlation or two distant thresholds are present, a smaller added value such as 0.1 or 0.25 becomes the better strategy.

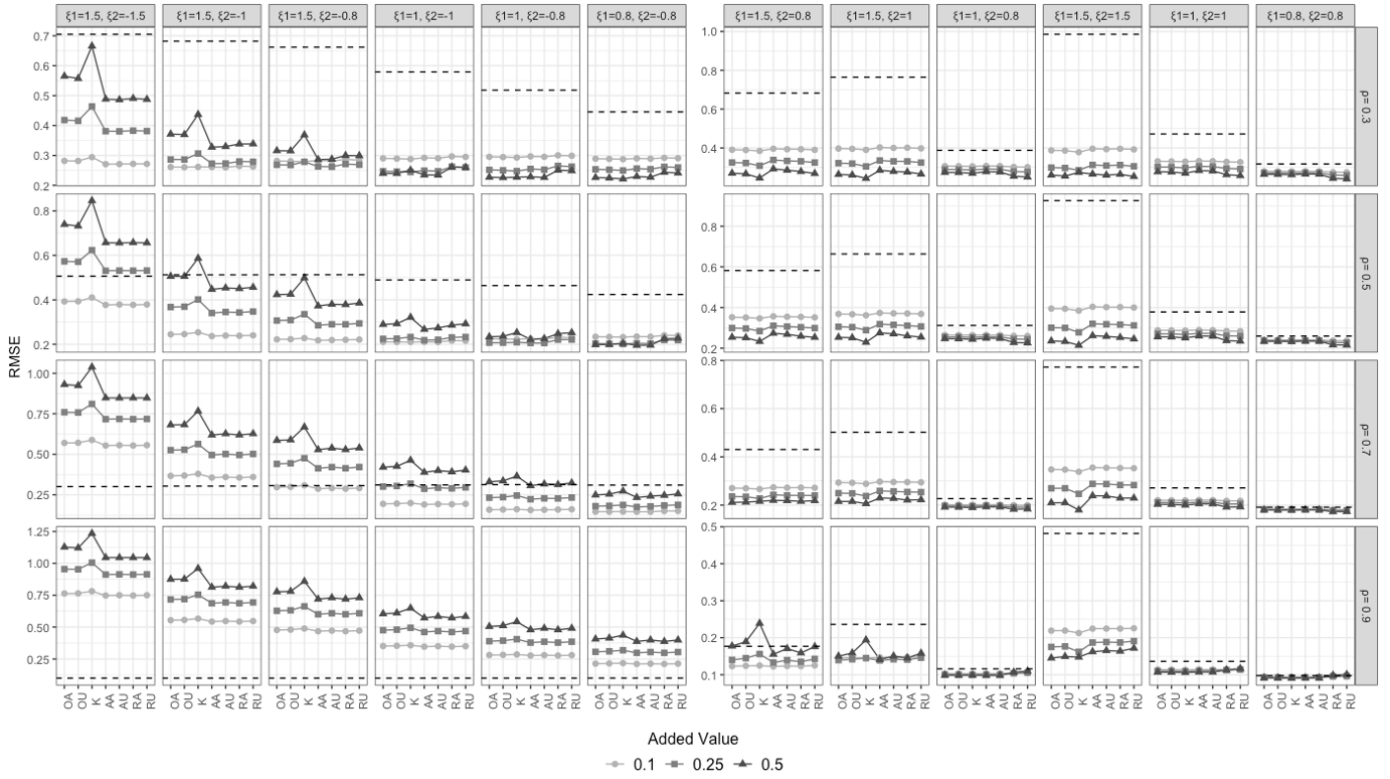


Figure 2: Root Mean Square Error (RMSE) for point estimates of the correlation ($N = 50$).

Note. Abbreviations for modifications: OA/OU = Only add to the zero cell and use adjusted/unadjusted thresholds in the second stage of estimation; K = Keep the marginal, for which the thresholds stay the same; AA/AU = Add to all cells when a zero is present in the table, and use adjusted/unadjusted thresholds in the second stage; RA/RU = Add to all cells regardless of the presence of zero, and use adjusted/unadjusted thresholds in the second stage. The dotted horizontal bar in each panel represents no correction. Y-scales vary with correlation sizes and threshold signs to better show differences within each panel.

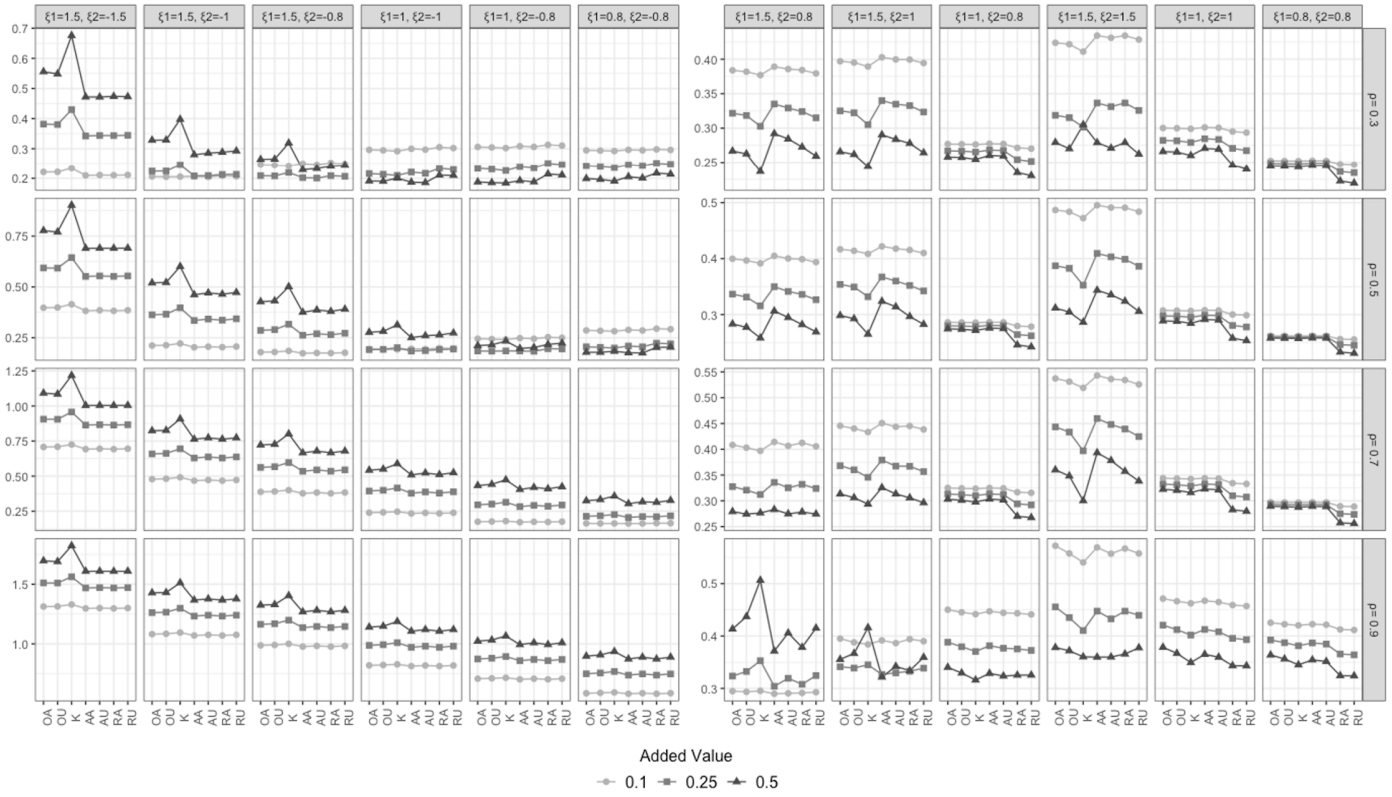


Figure 3: Mean Absolute Error (MAE) for point estimates of the correlation after Fisher's z-transformation ($N = 50$).

Note. The structure of this figure is the same as Figure 2. However, in this evaluation, no correction (i.e., added value 0) was not included.

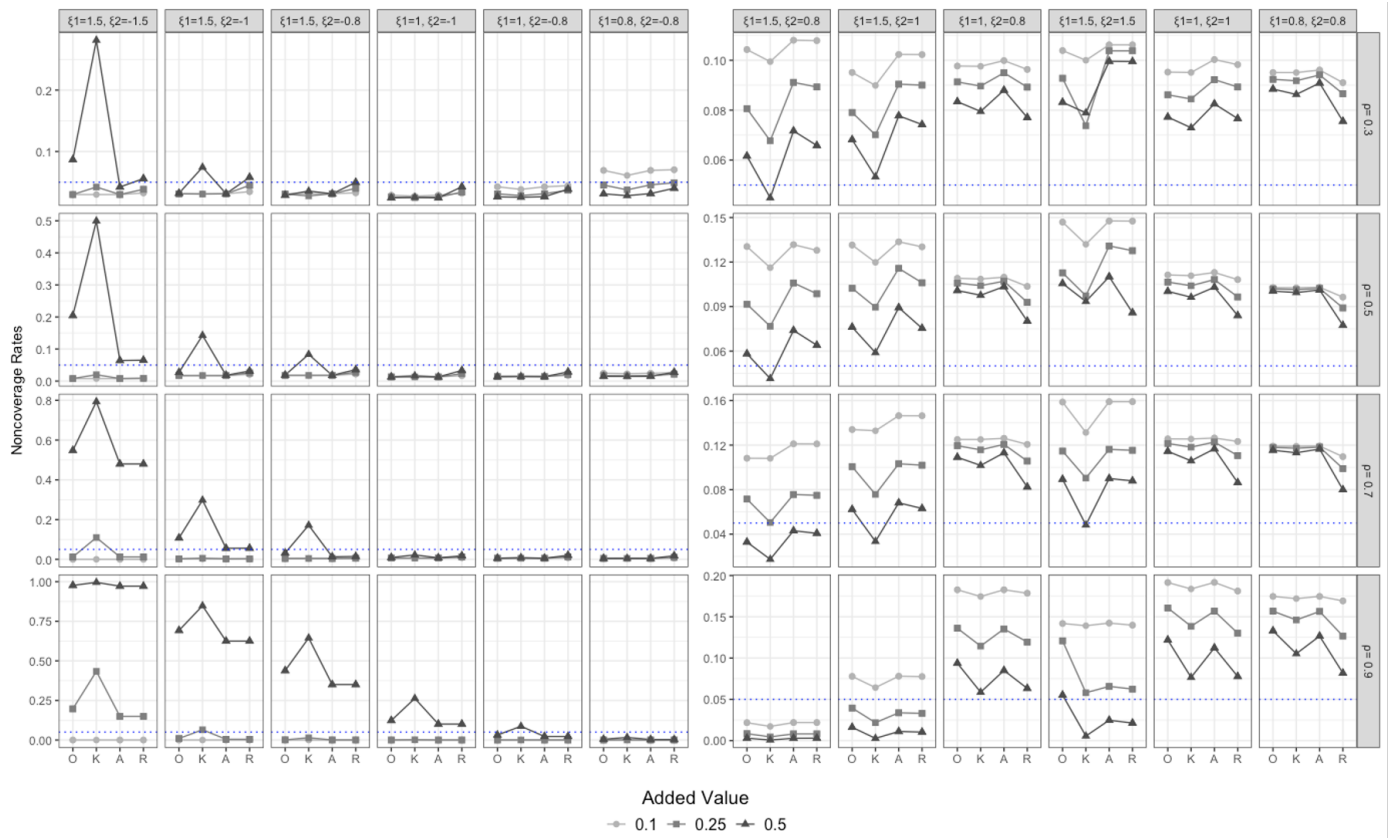


Figure 4: Noncoverage rate of the 95% Wald confidence interval of the correlation ($N = 50$).

Note. Abbreviations for modifications: O = Only add to the zero cell; K = Keep the marginal; A = Add to all cells when zero is present in the table; R = Add to all cells regardless of the presence of zero. The dotted horizontal bar in each panel indicates 0.05. Y-scales vary with correlation sizes and threshold signs.

3 Study 2: Multivariate Analysis

Given the common practice of using tetrachoric correlations in modeling with multiple variables rather than a single correlation, the second study was conducted to assess the different strategies' performance in estimating a correlation matrix.

3.1 Methods

Simulation Conditions The second simulation study used six binary variables. Data were first generated from a 6-variate normal distribution, and then discretized using thresholds. The sample size was set at 50, which produced the most notable results in the bivariate analysis.

To simplify the four correlations used in the bivariate simulation (0.3, 0.5, 0.7, and 0.9), we selected two representative values: 0.4 as the midpoint of 0.3 and 0.5 to represent moderate correlations, and 0.8 as the midpoint of 0.7 and 0.9 to represent high correlations. Based on these values, three correlation structures were considered: uniformly 0.4, uniformly 0.8, or mixed with two 3×3 diagonal blocks of 0.8 and an off-diagonal block of 0.4. Thresholds were also simplified using values from the bivariate simulation. Two sets of thresholds were considered: positively signed $(1.5, 1.0, 0.8, 1.5, 1.0, 0.8)$ and mixed signed $(-1.5, -1.0, -0.8, 1.5, 1.0, 0.8)$.

Crossing three sets of correlations and two sets of thresholds resulted in six generation conditions, each replicated 30,000 times. After the data generation, contingency tables were formed for each pair of variables, which were then used for estimating correlations.

Efficient Estimation of Correlations We evaluated case of no correction alongside 12 correction strategies, consistent with the bivariate simulation. However, we exclusively used adjusted thresholds in the second stage of the estimation, as their impact was found to be minimal in the bivariate simulation. Each of these correction strategies was applied to every contingency table for each pair of variables.

Unlike a traditional simulation where each replication from each simulation condition is analyzed, we only needed to estimate distinct tables among all replications across all conditions. Especially, given results from Study 1, the estimate of each distinct table was directly matched from our estimated results in Study 1 with a sample size of 50, to avoid redundant estimation and reduce computational cost. Then correlation matrices were constructed using these matched correlations. Because we had six variables in the analysis, we had a sample 6×6 correlation matrix including 15 different correlation estimates $\big(6 \times 5/2\big)$, along with diagonal elements being 1.

Evaluation Criteria We evaluated different strategies for handling zero cells based on two evaluation criteria: the percentage of positive definite correlation

matrices and the accuracy of estimation as measured by the average weighted squared error loss (a quadratic form loss).

First, the number of replications that yielded positive definite tetrachoric correlation matrices was counted. A positive definite matrix is necessary for many analyses, but having one does not mean the estimates are accurate. For example, adding larger constants to zero cells increases the chance of positive definiteness, but this can also move the estimates further away from the true matrix.

Second, we measured estimation error using the quadratic form loss. Let the sample correlation matrix be $\mathbf{R}$ and the population (true) correlation matrix be $\mathbf{P}$, then the quadratic form loss can be calculated as follows:

$$\mathrm{tr}\{(\mathbf{R}-\mathbf{P})\mathbf{P}^{-1}(\mathbf{R}-\mathbf{P})\mathbf{P}^{-1}\}. \tag{3}$$

Note that the matrix $\mathbf{P}$ is always positive definite, but the estimated matrix $\mathbf{R}$ may or may not be so. This criterion is the multivariate analogue of mean squared error. It can be seen as a variant of Stein's loss for covariance matrices a second-order Taylor expansion of Stein's loss around $\mathbf{P}$ yields the expression above. Unlike Stein's loss, however, it only requires $\mathbf{P}$ to be positive definite, not $\mathbf{R}$, which is useful here because pairwise tetrachoric estimates can produce a nonpositive definite $\mathbf{R}$.

3.2 Results

The Number of Zero-Frequency Cells We first conducted an analysis on the composition of generated tables across 30,000 replications. Each replication was examined for the presence of zero-frequency cells within the 15 pairwise contingency tables. Replications were classified based on the zero cells in these tables: if none of the 15 tables had any zero cells, the replication was categorized as "Without Zero-Frequency Cells." If there was at least one table with one zero cell, but none with two zero cells, it was categorized as "With a Zero-Frequency Cell." Finally, if at least one table included two or more zero cells, it was classified under "With Two or More Zero Cells (Removed)," and such replications were excluded from the analysis.

Table A4 in the Appendix shows that conditions with mixed correlations tend to have more instances of single zero-frequency cell than those with correlations consistently set at 0.4 or 0.8. With mixed thresholds, more replications within the condition include at least one table with zero-frequency cells, regardless of the correlation size, compared to positive thresholds. For each condition, the vast majority of the replications contain at least one table with a single zero-frequency cell. This suggests that our choice of simulation conditions is suitable for the evaluation of strategies to handle zero-frequency cells.

Positive Definiteness Results on positive definiteness are provided in Table A5 in the Appendix and Figure 5. The number of positive-definite matrices varies from 0 (0%) to 25,534 (85.11%), depending on the generation condition

and the correction method used. In this multivariate simulation, the added value also plays a major influence on the results. Positive definiteness is rare when no correction is performed on zero-frequency cells. More positive-definite matrices are observed in mixed threshold conditions compared to positive threshold conditions. Additionally, uniformly 0.4 and 0.8 correlations tend to produce more positive-definite matrices compared to mixed correlation conditions. Across all conditions, adding larger values tends to increase the count of positive-definite matrices. Specifically, adding an optimal value of 0.5 to the zero cell while keeping the marginal results in the highest count of positive-definite matrices.

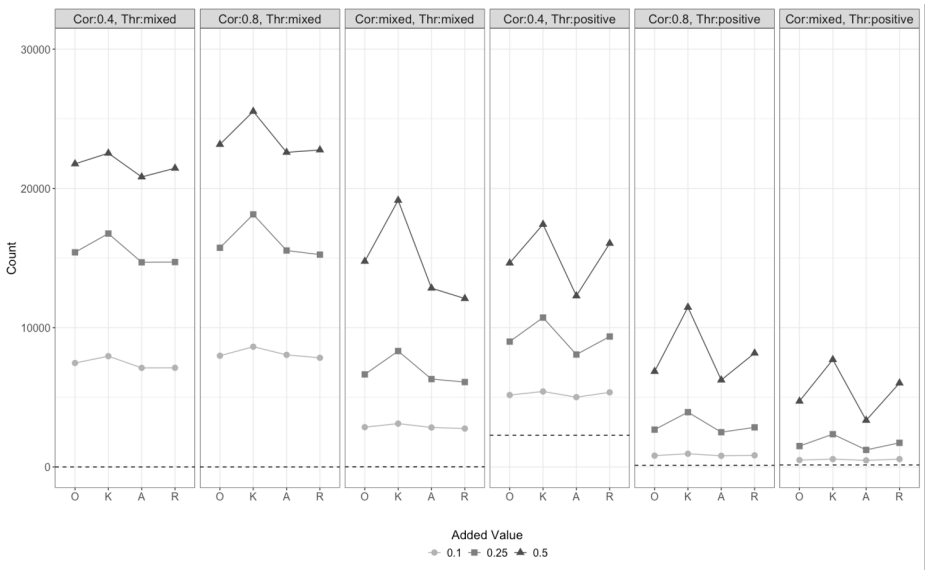


Figure 5: Number of positive definite matrices

Note. Abbreviations for modifications: O = Only add to the zero cell; K = Keep the marginal; A = Add to all cells when zero is present in the table; R = Add to all cells regardless of the presence of zero. The dotted horizontal bar in each panel represents no correction.

As the results suggest, adding 0.5 can make it more likely that the correlation matrix is positive definite, but a correction that achieves this by adding a larger constant may lead to estimates that deviate substantially from the population values. Thus, positive definiteness alone is not an adequate criterion for evaluating correction methods. It is also necessary to consider measures of deviation, such as quadratic form loss, which we introduce in the next section to assess how far the corrected estimates are from the true correlation structure.

Quadratic Form Loss Results regarding the quadratic form loss are shown in Table A6 in the Appendix and Figure 6. These results show that correlations being consistently 0.4 lead to smaller deviations compared to the other correlation conditions.

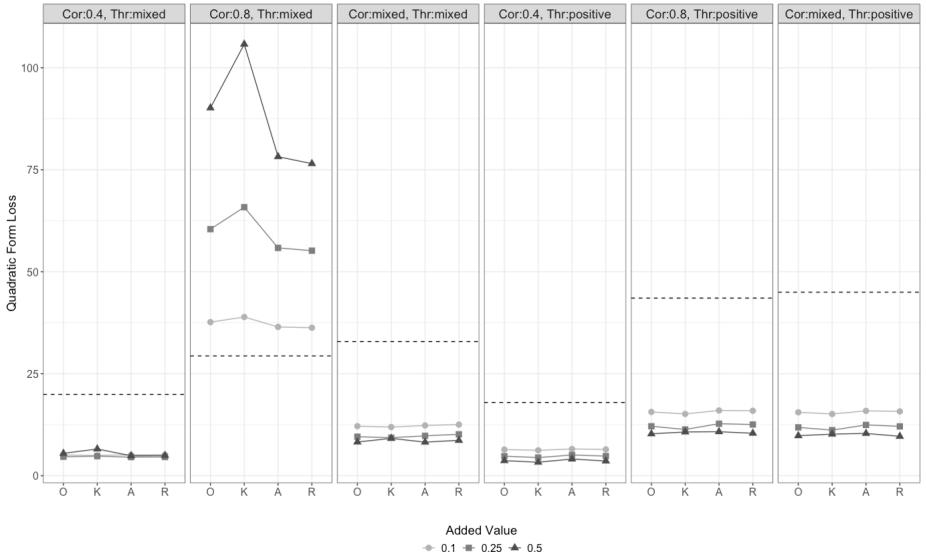


Figure 6: Quadratic form loss.

Note. This figure follows the same structure as Figure 5.

When thresholds are all positive, the quadratic form loss generally decreases as larger values are added, and not correcting for the zero cells results in notably higher deviations. With the optimal added value of 0.5, the manner in which it is added makes minimal differences.

When thresholds have mixed signs, the pattern differs by the size of correlations: when the correlations are 0.4, the quadratic form loss is lowest with the added value of 0.25; when the correlations are mixed between 0.4 and 0.8, adding 0.5 is the optimal added value; when the correlations are 0.8, the lowest loss occurs without zero-cell correction. In this third condition, adding larger values leads to higher deviations and the deviations are larger than those in the other panels. This is likely due to occasional sign flips with the correction: when the estimated correlation becomes negative, the distance from the true value (0.8) increases substantially, resulting in larger deviations. This will be discussed in further detail, with illustrative examples, in the Conclusion and Discussion section.

Implications of Multivariate Results and Their Relation to Bivariate Patterns Taken together, the positive definiteness and quadratic form loss results indicate that larger additions increase the likelihood of positive definiteness but move estimates farther from the true matrix. Thus, positive definiteness alone is therefore not an adequate criterion and it should be considered alongside accuracy measures such as the quadratic form loss. In our simulations, the strategy that produced the most positive definite matrices was not always the one that minimized quadratic form loss: larger corrections tended to favor positive definiteness, whereas smaller corrections often performed better when correlations were high and thresholds had mixed signs. This suggests that both criteria need to be considered together when evaluating correction strategies.

Overall, the multivariate results are generally consistent with the bivariate results. In both settings, the choice of added value primarily determines performance, while the manner of addition has little impact once the value chosen well. As in the bivariate patterns, higher correlations tend to favor smaller added values. With all-positive thresholds, adding 0.5 generally performs best, and leaving zero cells uncorrected performs poorly. With mixed-sign thresholds, at high correlation (e.g., 0.9 in the bivariate case or 0.8 in the multivariate case), no correction is preferable at high correlations (about 0.9 in the bivariate case and 0.8 in the multivariate case), whereas at lower correlations (around 0.4 in the multivariate case) smaller additions such as 0.1 or 0.25 work better.

4 Study 3: Confirmatory Factor Analysis

Since tetrachoric correlations are often used in fitting factor analysis models, in this simulation, different correction strategies for treating zero cells were evaluated according to their performance in estimating a confirmatory factor analysis (CFA) model.

4.1 Methods

Simulation Design This simulation involved four binary variables within a single-factor model. Data were generated from a 4-variate normal distribution with a sample size of 50 and then discretized using specific thresholds. Population loadings were consistently set at either 0.4 or 0.7. The threshold conditions for the four variables included either all positive thresholds or mixed-signed thresholds, with two positive and two negative values, and absolute values of 1.5, 1.0, and 0.8. This resulted in six distinct threshold conditions: all thresholds set to 1.5, all thresholds set to 1.0, all thresholds set to 0.8, mixed thresholds of 1.5 and -1.5, mixed thresholds of 1.0 and -1.0, and mixed thresholds of 0.8 and -0.8. By crossing the two loadings with the six sets of thresholds, twelve generation conditions were obtained, each replicated 1,000 times.

Computation With the generated data, contingency tables were created for each pair of variables. Consistent with Study 2, correlation and standard error

estimates for these tables were matched from the bivariate simulation results, and a 4×4 tetrachoric correlation matrix was produced for each replication. The resultant tetrachoric correlation matrix was further used to estimate the one-factor model with four variables through diagonally weighted least squares (DWLS). Specifically, we minimized the sum of squared differences between the estimated tetrachoric correlations and their model-implied values, with each difference weighted by the inverse of the estimated SE of the tetrachoric correlation. In the factor model, the factor variance was fixed at 1, the unique variances were constrained so that the latent continuous responses would have unit variances, and the factor loadings were constrained to lie between -1 and 1 to avoid Heywood cases. The `cfa` function from the `lavaan` package (Rosseel, 2012) was used.

Evaluation Criteria To evaluate the estimated loadings, we compared the mean of the four estimated loadings to the population loading. RMSE was used as the evaluation criterion and was computed after aligning the estimates. Before the estimated loadings could be properly evaluated, they were aligned across replications to address the potential sign indeterminacy in factor analysis. Note that in this CFA model all loadings can take the reversed signs to produce a statistically equivalent solution. To align the solutions across replications, for each replication, we computed the sum of squared differences between the estimated and true loadings, as well as for the true loadings with flipped signs. If the sum of squared differences was smaller for the sign-flipped true loadings, we flipped the signs of all estimated loadings for that replication.

4.2 Results

Figure 7 and Table A7 in the Appendix also show that the size of the RMSE is influenced by the added value. For all positive thresholds, adding larger values generally results in lower RMSE. Specifically, with the largest added value of 0.5, either keeping the marginals or adding to all cells regardless of zero cells often leads to the lowest RMSE. For mixed-sign thresholds, adding smaller values generally results in lower RMSE. With extreme thresholds of 1.5 regardless of the size of loadings, and thresholds of 1.0 with a loading of 0.7, the smallest added value of 0.1 results in the lowest RMSE. For thresholds of 1.0 with a loading of 0.4 and thresholds of 0.8, adding a medium (0.25) or large (0.5) value produces the lowest RMSE.

Relation to Bivariate Patterns These results are also generally consistent with the bivariate results. When thresholds are all positive, both show that larger added values lead to better performance, whether evaluated by RMSE of estimates or mean loadings. For mixed-sign thresholds, smaller added values work better in most cases. In less extreme situations, such as with smaller loadings or less extreme thresholds, added values like 0.25 or 0.5 give better results. These patterns are consistent with the bivariate results.

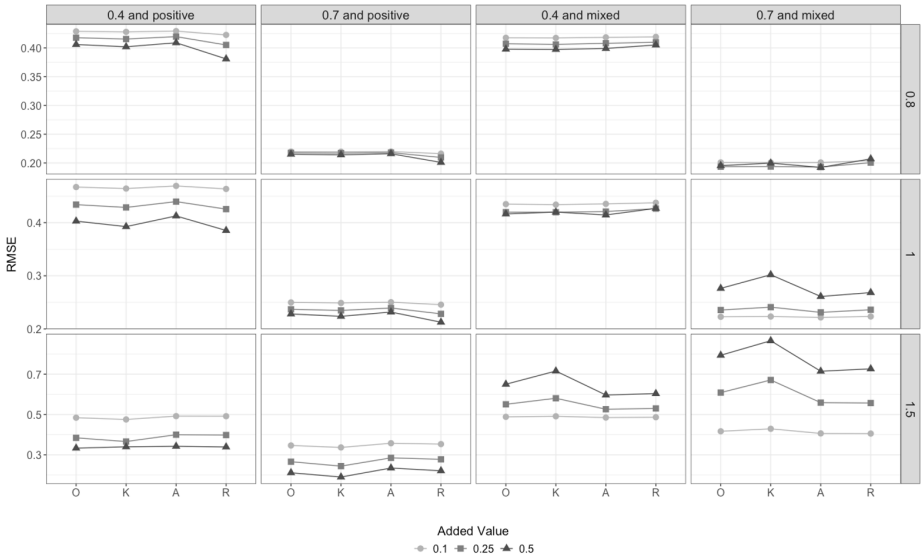


Figure 7: Root mean square error (RMSE) in the mean of loadings in the CFA simulation.

Note. This figure follows the same structure as Figure 6.

5 Conclusion & Discussion

5.1 Summary and Discussion of Results

This research addresses an important gap in the literature on tetrachoric correlation estimation by evaluating different strategies for treating zero-frequency cells. Practical approaches to zero-frequency cells are crucial for researchers estimating tetrachoric correlations. Despite available correction methods in statistical software for tetrachoric correlations, there are limited systematic evaluations on this issue, especially in multivariate contexts. Our simulation study makes a comprehensive effort to explore various correction methods of zero-frequency cell correction.

In the bivariate analysis, the choice of the added value had the largest impact on performance measures; for the optimal added value, the four different strategies made much less difference and the use of adjusted or unadjusted thresholds following correction had minimal effects. This optimal added value tended to be greater for a smaller correlation and similarly located thresholds but smaller for a more extreme correlation and more distantly located thresholds. Specifically, for most conditions with same-signed thresholds, adding a larger number, 0.5, while keeping the marginal produced the best results. For opposite-signed thresholds, the optimal added value remained at 0.5 for a small correlation and less distant thresholds but it decreased as the correlation and thresholds became more extreme; the best strategy became making no correction (or adding 0.1 if making

no correction was not feasible for the evaluation criterion) for a correlation of 0.9. Table 1 summarizes the main patterns in the bivariate analysis.

Table 1: Summary of optimal added-value patterns in bivariate simulation

Scenario	Pattern	Best added value
Overall (across conditions)	Choice of added value had the largest impact on performance. When the added value was near-optimal, differences among strategies were small, and using adjusted versus unadjusted thresholds after correction had minimal effects.	–
Larger optimal added value	More likely with smaller correlations and thresholds that are closer (similarly located).	Higher
Smaller optimal added value	More likely with more extreme correlations and thresholds that are farther apart (more distantly located).	Lower
Same-signed thresholds (most conditions)	Adding a larger number while keeping marginals fixed produced the best results.	0.5
Opposite-signed thresholds (small correlation, less distant thresholds)	Optimal added value remained high.	0.5
Opposite-signed thresholds (more extreme correlation and/or more distant thresholds)	Optimal added value decreased as correlation and thresholds became more extreme.	Decreasing trend
Opposite-signed thresholds (very high correlation)	Best became making no correction; if not feasible for the evaluation criterion (fisher's z), use a minimal correction.	0 (or 0.1 if needed)

Smaller optimal added values under extreme thresholds and/or large correlations likely reflect how sensitive tetrachoric estimates are to small corrections when zero cells occur. In these settings, adding a value to a zero cell can noticeably shift the estimate, and in some cases even reverse its sign. Figure 8 illustrates two examples of this sign-flip behavior. In the top panel of Figure 8, under a simulation condition with population correlation 0.3 and thresholds ± 1.5, an example table with observed counts $(47, 2, 1, 0)$ contains a zero cell because one theoretical cell probability is extremely small. For this table, the MLE without correction is 1, although software may report a value close to 1 due to imper-

fect convergence. When 0.5 is added only to the zero cell, the estimate becomes -0.5667. This example shows that, with opposite-signed thresholds, a correction can substantially change the estimate and even flip its sign, which explains why the optimal added value tends to be smaller in more extreme settings.

Negative-correlation cases were not included in our simulation condition because they can be re-expressed within our design by flipping the threshold sign pattern: a negative correlation with same-signed thresholds corresponds to a positive correlation with mixed-signed thresholds, and a negative correlation with mixed-signed thresholds corresponds to a positive correlation with same-signed thresholds. For illustration, the bottom panel of Figure 8 shows the corresponding case with both thresholds negative and demonstrates that the sign-flip behavior can also occur in the opposite direction: in the top panel the correction induces a sign flip, whereas in the bottom panel the uncorrected estimate flips sign and the correction yields an estimate with the same sign as the population value.

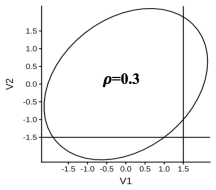
Given that a simulation study can only consider a limited number of added values, it is highly likely that the actual optimal added value lies between 0 and 0.1 in this situation. Our findings show that in situations where adding a number is advantageous, such as with opposite-signed thresholds found in Savalei (2011) and Yang and Weng (2024), the optimal value to add can vary. This adds more detailed information compared to the previous studies.

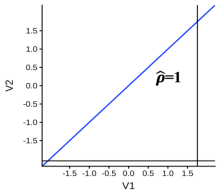
In multivariate simulation, the combination of correlations and thresholds led to different outcomes. Adding a larger number often yielded better results, while not correcting for the zero cell consistently resulted in poor results. Within these added values, keeping the marginals was generally the most effective strategy for achieving positive definiteness. However, for high correlations with mixed-signed thresholds, not correcting for zero cells or adding a smaller value often yielded better results in terms of the quadratic form loss. The results show that methods improving positive definiteness do not always give the most accurate estimates, so both aspects should be considered when choosing a correction strategy.

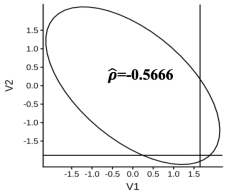
In CFA simulation, the optimal added value depended on correlation and threshold sizes. Smaller added values produced better results with extreme mixed-signed thresholds, while larger values were more effective for positive thresholds.

Although no single approach performed best across all studies, several consistent patterns emerged across the bivariate, multivariate, and CFA scenarios. First, when thresholds are of the same sign, adding 0.5 tends to work well across settings and can be a reasonable choice in many situations. Second, when thresholds have opposite signs, smaller additions such as 0.1 or 0.25 may work better, while in rare cases of very high correlations leaving the zero cells uncorrected can perform similarly. Third, the specific manner the value is added appears less important than the size of the addition itself. Finally, the effect of correction should be evaluated from multiple perspectives: larger corrections increase the likelihood of obtaining a positive definite matrix but do not necessarily improve estimation accuracy, as positive definiteness becomes more likely as a consequence of larger corrections.

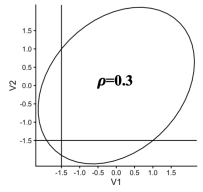
True ρ						$\widehat{\rho}$ (No Modification)						$\widehat{\rho}$ (With Modification)					
Probabilities			Expected Table			Probabilities			Observed Table			Probabilities			Modified Table		
	0	1		0	1		0	1		0	1		0	1		0	1
1	0.867	0.066	1	43.35	3.30	1	0.94	0.04	1	47	2	1	0.93	0.04	1	47	2
0	0.066	0.001	0	3.30	0.05	0	0.02	0	0	1	0	0	0.02	0.01	0	1	0.5

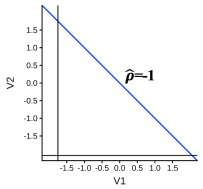






True ρ						$\widehat{\rho}$ (No Modification)						$\widehat{\rho}$ (With Modification)					
Probabilities			Expected Table			Probabilities			Observed Table			Probabilities			Modified Table		
	0	1		0	1		0	1		0	1		0	1		0	1
1	0.06	0.88	1	2.77	43.89	1	0.04	0.94	1	2	47	1	0.04	0.93	1	2	47
0	0.01	0.06	0	0.57	2.77	0	0.00	0.02	0	0	1	0	0.01	0.02	0	0.5	1





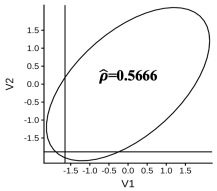


Figure 8: Examples of sign flips in the estimated tetrachoric correlation with and without correction

Note. Each panel shows the bivariate normal cell probabilities computed with the corresponding correlation and threshold values, together with an example 2×2 table and the resulting tetrachoric correlation estimates. For the estimate without correction (with a zero cell), $\widehat{\rho} = \pm 1$; however, different software packages may give slightly different results due to lack of perfect convergence.

We would like to note that our simulation focused on the estimation of tetrachoric correlations, whereas polychoric correlations involve more than two categories and can behave differently as discussed in the Introduction. Accordingly, our findings may not generalize to polychoric settings, consistent with Savalei's (2011) conclusions.

5.2 Contributions

The current study makes several contributions to the literature on estimating tetrachoric correlations in the presence of zero-frequency cells. Firstly, we provided a more comprehensive understanding of the zero-cell corrections in the tetrachoric correlation context. Prior work, such as Savalei (2011), primarily compared adding 0.5 to not adding it. Our study included more correction methods and revealed a potential optimal value to add between 0 and 0.5. We revealed unexplored distinctions in different ways of adding the number, including the issue of using adjusted and unadjusted thresholds in the two-stage procedure.

Secondly, these different correction methods were examined under more realistic data conditions and practical scenarios. Specifically, we included a sample size of 50 to better depict smaller sample situations where zero-frequency cells are more prevalent. Additionally, we considered multivariate and confirmatory factor analysis scenarios, reflecting the practical use of tetrachoric correlations in constructing correlation matrices for various analyses. In this way, we illustrate how different correction strategies affect not only individual correlation estimates but entire models.

Thirdly, this study provided a more efficient way of conducting simulation studies so that researchers can significantly reduce the number of estimations. Tetrachoric correlations are known to be computationally intensive (Zhang et al., 2022). To alleviate this problem, we took advantage of 2×2 discrete nature of tetrachoric correlations. We did this by simplifying the sets of the simulation condition and reducing the number of estimations by setting the prototype of similar tables. As this computation issue can get worse in simulation studies that involve many replications, our approach to make the simulation efficient can serve as a good example of simulations with tetrachoric correlations. It also has potential to be extended as a more general framework for discrete data simulations in broader contexts.

Lastly, we reduced simulation error by deriving the sampling distribution from all possible tables under each condition, not relying on a limited number of replications. This framework avoids error introduced by random sampling and produces more precise simulation results. The influence of sampling error is especially evident in discrete data problems such as 2×2 tables, where the limited number of replications can exaggerate variability. Therefore, this approach is useful for generating more accurate and reliable simulation results and helps to overcome fundamental limitations of simulation studies involving discrete data.

5.3 Limitations and Future Directions

Despite these contributions, there is a need for additional exploration in future research. Firstly, our analysis is limited in assessing potentially effective ways to correct zero cells due to simulation design constraints. Future research could consider a wider range of values such as those smaller than 0.1. For multivariate simulations, applying different correction methods to each pair of variables could potentially provide useful insights, while our study applied the same correction across the variables to see the general trend. Secondly, exploring varied correction methods for each variable in multivariate simulations and investigating alternative approaches, such as collapsing categories (DiStefano, Shi, & Morgan, 2021) instead of adding a small number, are also possible directions. We anticipate future research to investigate more meticulous approaches for addressing zero-frequency cells.

Author Notes

Parts of the results of this paper were presented at the 2024 Annual Meeting of the Society of Multivariate Experimental Psychology (Choi & Wu, 2025).

References

Brown, M. B., & Benedetti, J. K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, *42*(3), 347–355. doi: https://doi.org/10.1007/bf02293655

Choi, J., & Wu, H. (2025). On zero-count correction strategies in tetrachoric correlation estimation (abstract). *Multivariate Behavioral Research*, *60*(1), 3–4. doi: https://doi.org/10.1080/00273171.2024.2442249

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in Psychology*, *9*, 580. doi: https://doi.org/10.3389/fpsyg.2018.00580

DiStefano, C., Shi, D., & Morgan, G. B. (2021). Collapsing categories is often more advantageous than modeling sparse data: Investigations in the cfa framework. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 237–249. doi: https://doi.org/10.1080/10705511.2020.1803073

Fox, J. (2022). *polycor: Polychoric and polyserial correlations.* Retrieved from `https://CRAN.R-project.org/package=polycor` (R package version 0.8-1)

Golino, H., & Christensen, A. P. (2025). *Eganet: Exploratory graph analysis – a framework for estimating the number of dimensions in multivariate data using network psychometrics.* Retrieved from `https://r-ega.net` (R package version 2.0.3)

Muthén, L. K., & Muthén, B. O. (2017). *Mplus.* (Version 8)

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. doi: https://doi.org/10.1007/bf02296207

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Retrieved from `https://www.R-project.org` (Version 4.2.1)

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research.* Retrieved from `https://CRAN.R-project.org/package=psych` (R package version 2.3.9)

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: https://doi.org/10.18637/jss.v048.i02

Savalei, V. (2011). What to do about zero frequency cells when estimating polychoric correlations. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 253–273. doi: https://doi.org/10.1080/10705511.2011.557339

Yang, T.-R., & Weng, L.-J. (2024). Revisiting Savalei's (2011) research on remediating zero-frequency cells in estimating polychoric correlations: A data distribution perspective. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(1), 81–96. doi: https://doi.org/10.1080/10705511.2023.2220919

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217–235. doi: https://doi.org/10.2307/2983604

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modeling with correlation matrices for ordinal and continuous data. *The British Journal of Mathematical and Statistical Psychology*, *64*(1), 107-133. doi: https://doi.org/10.1348/000711010x497442

Zhang, G., Trichtinger, L. A., Lee, D., & Jiang, G. (2022). PolychoricRM: A computationally efficient R function for estimating polychoric correlations and their asymptotic covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 310–320. doi: https://doi.org/10.1080/10705511.2021.1929996

Appendix

Table A1: The Composition of Sampling Distribution of 2×2 Tables for $N = 50$ (Study 1)

ξ_1		1.5	1.5	1.5	1	1	0.8	1.5	1.5	1	1.5	1	0.8
ξ_2		-1.5	-1	-0.8	-1	-0.8	-0.8	0.8	1	0.8	1.5	1	0.8
ρ	Zero Cells	Probabilities											
0.3	0	0.046	0.147	0.213	0.404	0.535	0.674	0.648	0.603	0.943	0.383	0.898	0.976
	1	0.892	0.821	0.756	0.595	0.465	0.326	0.320	0.365	0.057	0.554	0.102	0.024
	2 or 3	0.062	0.032	0.032	0.000	0.000	0.000	0.032	0.032	0.000	0.063	0.000	0.000
0.5	0	0.008	0.041	0.070	0.173	0.271	0.403	0.654	0.664	0.967	0.504	0.948	0.989
	1	0.930	0.928	0.898	0.827	0.729	0.597	0.315	0.304	0.033	0.432	0.052	0.011
	2 or 3	0.062	0.032	0.032	0.000	0.000	0.000	0.032	0.032	0.000	0.065	0.000	0.000
0.7	0	0.000	0.003	0.006	0.025	0.054	0.111	0.504	0.595	0.944	0.559	0.947	0.981
	1	0.938	0.966	0.962	0.974	0.945	0.889	0.465	0.373	0.056	0.368	0.052	0.019
	2 or 3	0.062	0.032	0.032	0.000	0.000	0.000	0.032	0.032	0.000	0.073	0.001	0.000
0.9	0	0.000	0.000	0.000	0.000	0.000	0.000	0.105	0.224	0.713	0.415	0.789	0.864
	1	0.938	0.968	0.968	1.000	1.000	1.000	0.863	0.741	0.282	0.444	0.200	0.131
	2 or 3	0.062	0.032	0.032	0.000	0.000	0.000	0.032	0.036	0.004	0.141	0.011	0.004

Note. ξ_1 and ξ_2 are thresholds for the first and second variables, respectively. Probabilities were rounded to the third decimal place.

Table A2: The Composition of Sampling Distribution of 2×2 Tables for $N = 100$ (Study 1)

ξ_1		1.5	1.5	1.5	1	1	0.8	1.5	1.5	1	1.5	1	0.8
ξ_2		-1.5	-1	-0.8	-1	-0.8	-0.8	0.8	1	0.8	1.5	1	0.8
ρ	Zero Cells	Probabilities											
0.3	0	0.095	0.281	0.393	0.641	0.778	0.886	0.917	0.881	0.988	0.674	0.983	0.988
	1	0.902	0.715	0.603	0.354	0.215	0.105	0.077	0.114	0.002	0.322	0.009	0.000
	2 or 3	0.002	0.001	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.002	0.000	0.000
	Removed	0.001	0.003	0.003	0.005	0.007	0.009	0.005	0.004	0.010	0.002	0.008	0.012
0.5	0	0.017	0.082	0.139	0.313	0.465	0.639	0.921	0.931	0.990	0.829	0.990	0.988
	1	0.981	0.916	0.858	0.684	0.531	0.355	0.073	0.065	0.000	0.167	0.001	0.000
	2 or 3	0.002	0.001	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.002	0.000	0.000
	Removed	0.000	0.001	0.002	0.003	0.004	0.006	0.005	0.003	0.010	0.002	0.009	0.012
0.7	0	0.000	0.005	0.012	0.048	0.104	0.206	0.785	0.875	0.989	0.904	0.992	0.989
	1	0.997	0.993	0.985	0.949	0.893	0.790	0.210	0.121	0.002	0.091	0.000	0.000
	2 or 3	0.002	0.001	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.002	0.000	0.000
	Removed	0.001	0.001	0.002	0.003	0.003	0.004	0.004	0.003	0.009	0.003	0.008	0.011
0.9	0	0.000	0.000	0.000	0.000	0.000	0.000	0.204	0.413	0.920	0.801	0.972	0.984
	1	0.998	0.999	0.998	0.999	0.999	0.998	0.793	0.584	0.074	0.187	0.023	0.009
	2 or 3	0.002	0.001	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.011	0.000	0.000
	Removed	0.000	0.000	0.001	0.001	0.001	0.002	0.002	0.002	0.006	0.001	0.005	0.007

Note. The structure of this table is the same as Table A1. "Removed" indicates tables excluded from analyses due to their small probability in the sampling distribution. "2 or 3" was also removed from the analysis due to the number of zero cells.

Table A3: The Composition of Sampling Distribution of 2×2 Tables for $N = 200$ (Study 1)

ξ_1		1.5	1.5	1.5	1	1	0.8	1.5	1.5	1	1.5	1	0.8
ξ_2		-1.5	-1	-0.8	-1	-0.8	-0.8	0.8	1	0.8	1.5	1	0.8
ρ	Zero Cells	Probabilities											
0.3	0	0.180	0.482	0.628	0.863	0.937	0.967	0.983	0.978	0.972	0.892	0.976	0.966
	1	0.817	0.512	0.364	0.124	0.045	0.009	0.003	0.011	0.000	0.102	0.000	0.000
	2 or 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Removed	0.003	0.006	0.008	0.013	0.018	0.024	0.014	0.011	0.028	0.006	0.024	0.034
0.5	0	0.033	0.156	0.258	0.524	0.707	0.859	0.984	0.987	0.972	0.969	0.976	0.967
	1	0.965	0.840	0.737	0.467	0.281	0.125	0.002	0.001	0.000	0.024	0.000	0.000
	2 or 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Removed	0.002	0.004	0.005	0.009	0.012	0.016	0.014	0.012	0.028	0.007	0.024	0.033
0.7	0	0.000	0.009	0.024	0.094	0.196	0.367	0.947	0.977	0.975	0.990	0.978	0.970
	1	0.998	0.988	0.973	0.901	0.797	0.624	0.042	0.012	0.000	0.003	0.000	0.000
	2 or 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Removed	0.002	0.003	0.003	0.005	0.007	0.009	0.011	0.011	0.025	0.007	0.022	0.030
0.9	0	0.000	0.000	0.000	0.000	0.000	0.000	0.366	0.653	0.979	0.976	0.985	0.979
	1	0.999	0.999	0.999	0.998	0.998	0.996	0.629	0.341	0.004	0.019	0.000	0.000
	2 or 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Removed	0.001	0.001	0.001	0.002	0.002	0.004	0.005	0.006	0.017	0.005	0.015	0.021

Note. The structure of this table is the same as Table A2. "Removed" indicates tables excluded from analyses due to their small probability in the sampling distribution. "2 or 3" was also removed from the analysis due to the number of zero cells. The removed cases are those with very small probabilities. As sample size increases, most replications concentrate on 2×2 tables close to the expected table, so many other possible tables become unlikely. As a result, more tables fall into the very low probability range and are removed, which is why the larger-sample condition shown in this table ends up with a larger total probability of removed tables.

Table A4: The Composition of the Simulated Dataset (Study 2)

Correlation	Thresholds	No Zero Cells	One Zero Cell	Two or More Zero Cells (Removed)
0.4	positive	4207	23866	1927
	mixed	0	28239	1761
0.8	positive	409	26683	2908
	mixed	0	27947	2053
mixed	positive	715	27215	2070
	mixed	19	28004	1977

Note. The total number of replications is 30,000 for each condition.

Table A5: The Proportion of Positive-Definite Matrices (Study 2)

Added Value	Way of Adding	Cor 0.4 & Mixed Thresholds	Cor 0.8 & Mixed Thresholds	Cor Mixed & Mixed Thresholds	Cor 0.4 & Positive Thresholds	Cor 0.8 & Positive Thresholds	Cor Mixed & Positive Thresholds
0.00	–	0.000	0.000	0.000	0.076	0.004	0.005
0.10	Only to the Zero Cell	0.249	0.266	0.095	0.172	0.027	0.017
	Keep Marginals	0.265	0.288	0.104	0.181	0.032	0.019
	Add to All	0.237	0.268	0.094	0.167	0.027	0.016
	Add to All Regardless	0.237	0.261	0.092	0.179	0.028	0.019
0.25	Only to the Zero Cell	0.514	0.525	0.222	0.300	0.089	0.050
	Keep Marginals	0.559	0.605	0.277	0.358	0.131	0.078
	Add to All	0.490	0.518	0.210	0.269	0.083	0.041
	Add to All Regardless	0.490	0.508	0.203	0.312	0.095	0.058
0.50	Only to the Zero Cell	0.726	0.772	0.492	0.488	0.229	0.157
	Keep Marginals	0.751	0.851	0.638	0.581	0.382	0.257
	Add to All	0.694	0.753	0.428	0.410	0.208	0.112
	Add to All Regardless	0.715	0.759	0.403	0.535	0.272	0.201

Table A6: Quadratic Form Loss (Study 2)

Added Value	Way of Adding	Cor 0.4 & Mixed Thresholds	Cor 0.8 & Mixed Thresholds	Cor Mixed & Mixed Thresholds	Cor 0.4 & Positive Thresholds	Cor 0.8 & Positive Thresholds	Cor Mixed & Positive Thresholds
0.00	–	19.944	29.372	32.886	17.955	43.544	44.983
0.10	Only to the Zero Cell	5.047	37.650	12.150	6.402	15.629	15.550
	Keep Marginals	4.987	38.910	11.922	6.233	15.136	15.121
	Add to All	5.095	36.470	12.311	6.541	15.992	15.891
	Add to All Regardless	5.107	36.280	12.530	6.430	15.920	15.765
0.25	Only to the Zero Cell	4.646	60.434	9.554	4.787	12.121	11.843
	Keep Marginals	4.793	65.822	9.301	4.439	11.327	11.161
	Add to All	4.557	55.836	9.772	5.073	12.743	12.448
	Add to All Regardless	4.592	55.154	10.144	4.807	12.551	12.092
0.50	Only to the Zero Cell	5.472	90.158	8.249	3.687	10.260	9.799
	Keep Marginals	6.553	105.776	9.148	3.306	10.715	10.187
	Add to All	4.876	78.200	8.221	4.088	10.780	10.365
	Add to All Regardless	4.954	76.497	8.667	3.605	10.401	9.648

Note. Values are rounded to the third decimal place.

Table A7: Root Mean Square Error (RMSE) in the Mean of Loadings (Study 3)

		Loadings of 0.4 & Positive Thresholds			Loadings of 0.7 & Positive Thresholds			Loadings of 0.4 & Mixed Thresholds			Loadings of 0.7 & Mixed Thresholds		
Added Value	Way of Adding	0.8	1	1.5	0.8	1	1.5	0.8	1	1.5	0.8	1	1.5
0.1	Only to the zero cell	0.429	0.467	0.484	0.220	0.250	0.346	0.418	0.435	0.488	0.201	0.223	0.417
	Keep marginals	0.428	0.464	0.475	0.220	0.249	0.337	0.417	0.434	0.491	0.201	0.223	0.429
	Add to all	0.429	0.469	0.492	0.220	0.250	0.357	0.418	0.435	0.485	0.201	0.222	0.406
	Add to all regardless	0.423	0.464	0.492	0.216	0.246	0.353	0.419	0.437	0.486	0.204	0.223	0.405
0.25	Only to the zero cell	0.418	0.434	0.384	0.218	0.237	0.266	0.407	0.420	0.551	0.194	0.236	0.609
	Keep marginals	0.415	0.429	0.366	0.217	0.235	0.244	0.406	0.420	0.581	0.194	0.241	0.671
	Add to all	0.420	0.440	0.400	0.218	0.239	0.285	0.408	0.421	0.526	0.193	0.231	0.559
	Add to all regardless	0.405	0.426	0.398	0.210	0.228	0.278	0.410	0.427	0.530	0.201	0.236	0.557
0.5	Only to the zero cell	0.406	0.403	0.333	0.215	0.228	0.211	0.398	0.416	0.650	0.195	0.276	0.794
	Keep marginals	0.402	0.393	0.340	0.214	0.224	0.190	0.397	0.420	0.716	0.200	0.302	0.866
	Add to all	0.409	0.413	0.342	0.216	0.232	0.234	0.399	0.415	0.597	0.192	0.261	0.715
	Add to all regardless	0.381	0.385	0.339	0.201	0.213	0.221	0.405	0.427	0.604	0.207	0.268	0.727

Note. Values are rounded to the third decimal place.

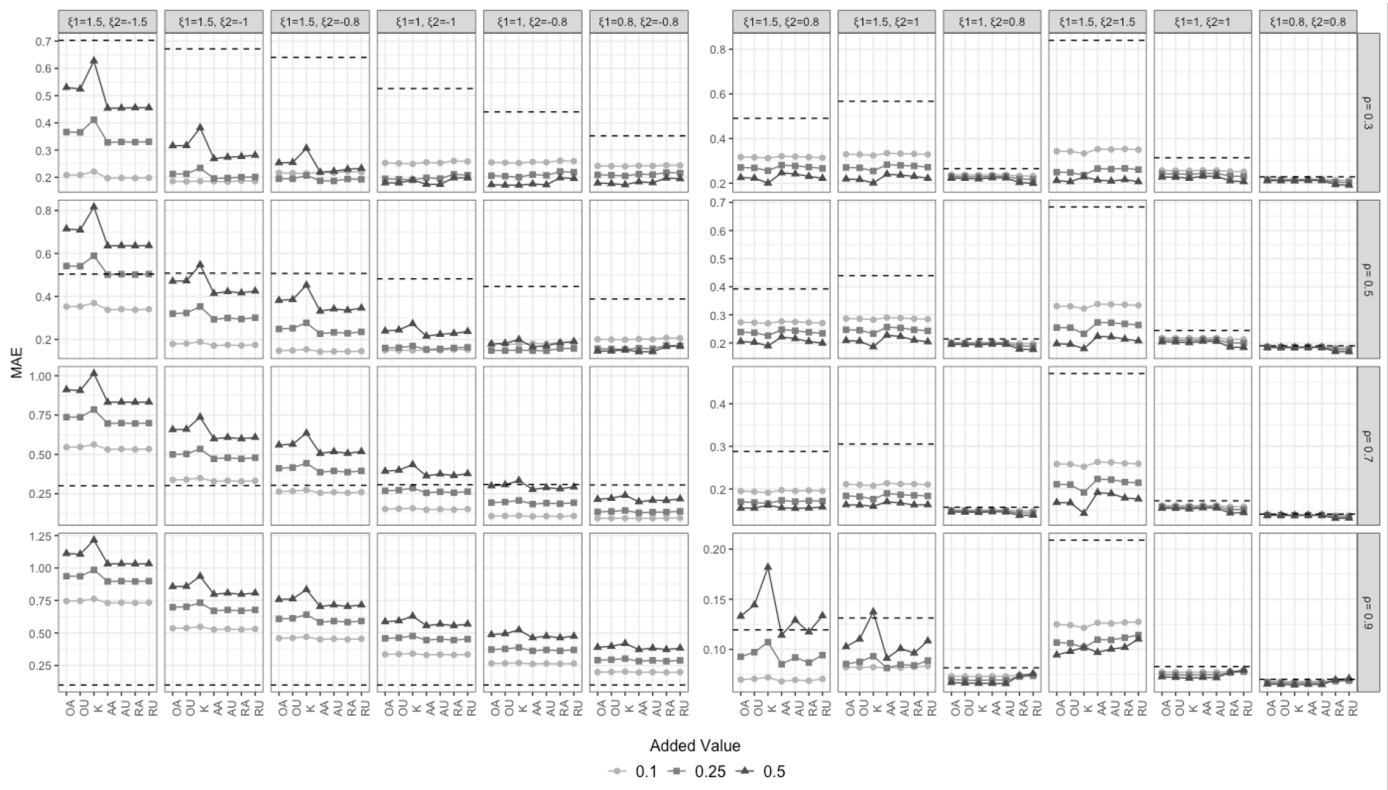


Figure A1: Mean Absolute Error (MAE) for Point Estimates (N=50)

Note. The structure of this figure is the same as Figure 2 in the paper.

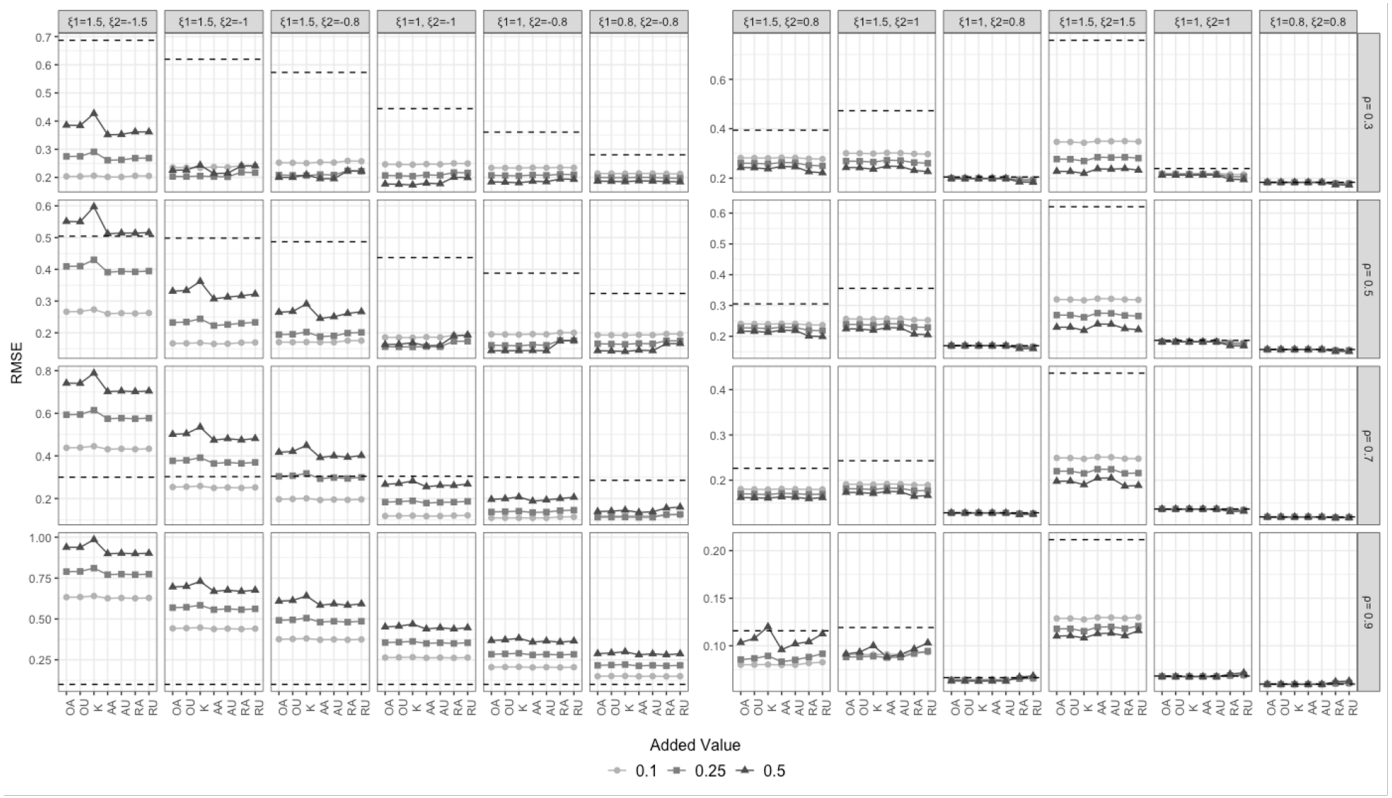


Figure A2: Root Mean Square Error (RMSE) for Point Estimates (N=100)

Note. The structure of this figure is the same as Figure 2 in the paper.

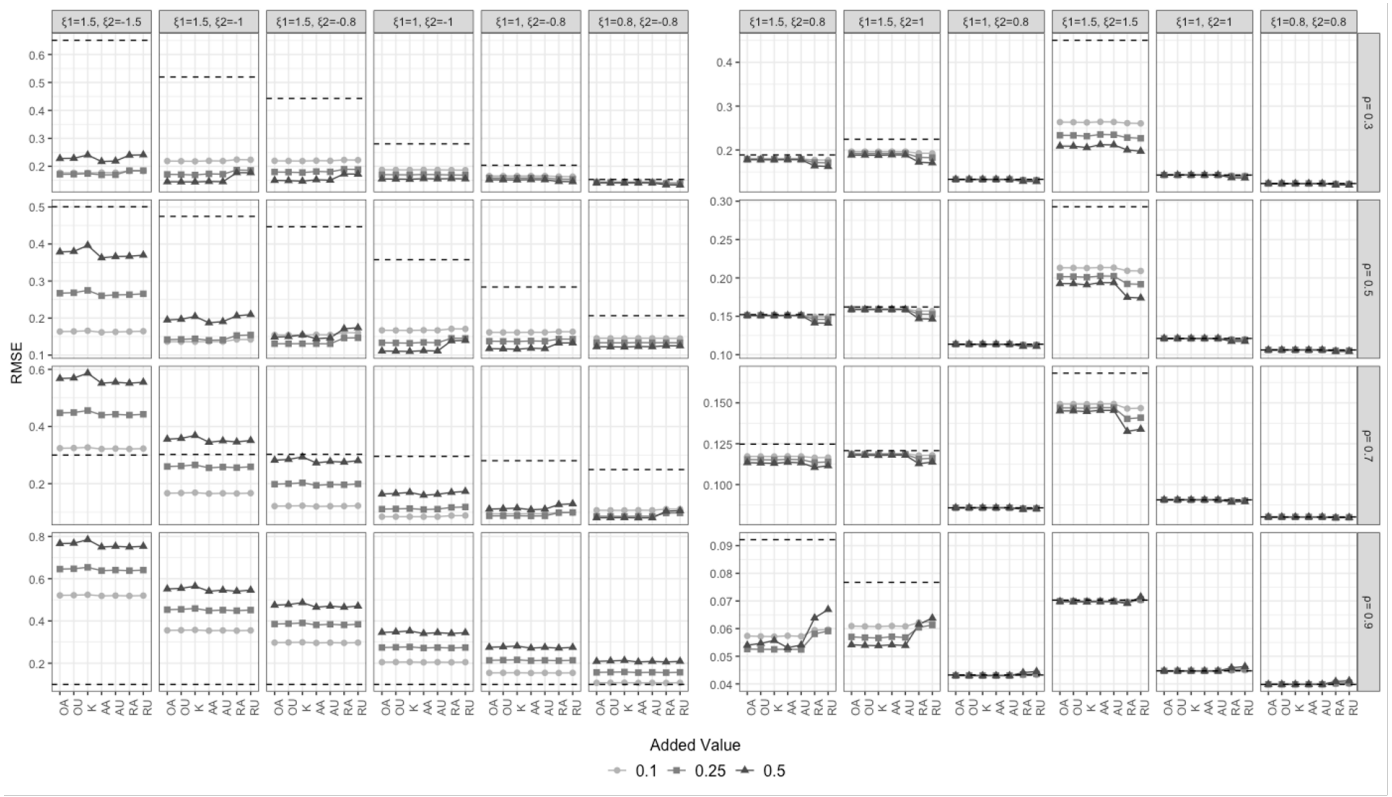


Figure A3: Root Mean Square Error (RMSE) for Point Estimates (N=200)

Note. The structure of this figure is the same as Figure 2 in the paper.

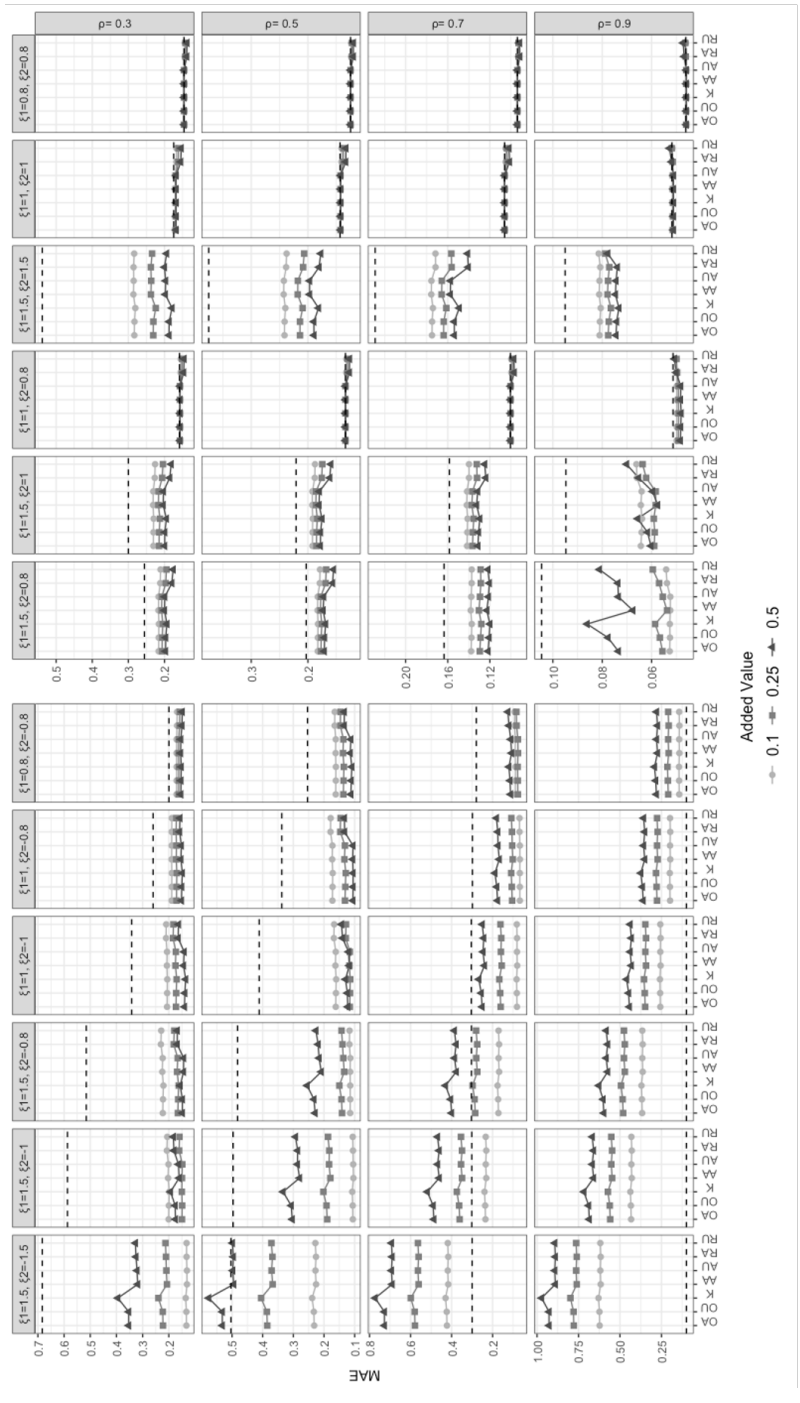


Figure A4: Mean Absolute Error (MAE) for Point Estimates (N=100)

Note. The structure of this figure is the same as Figure 2 in the paper.

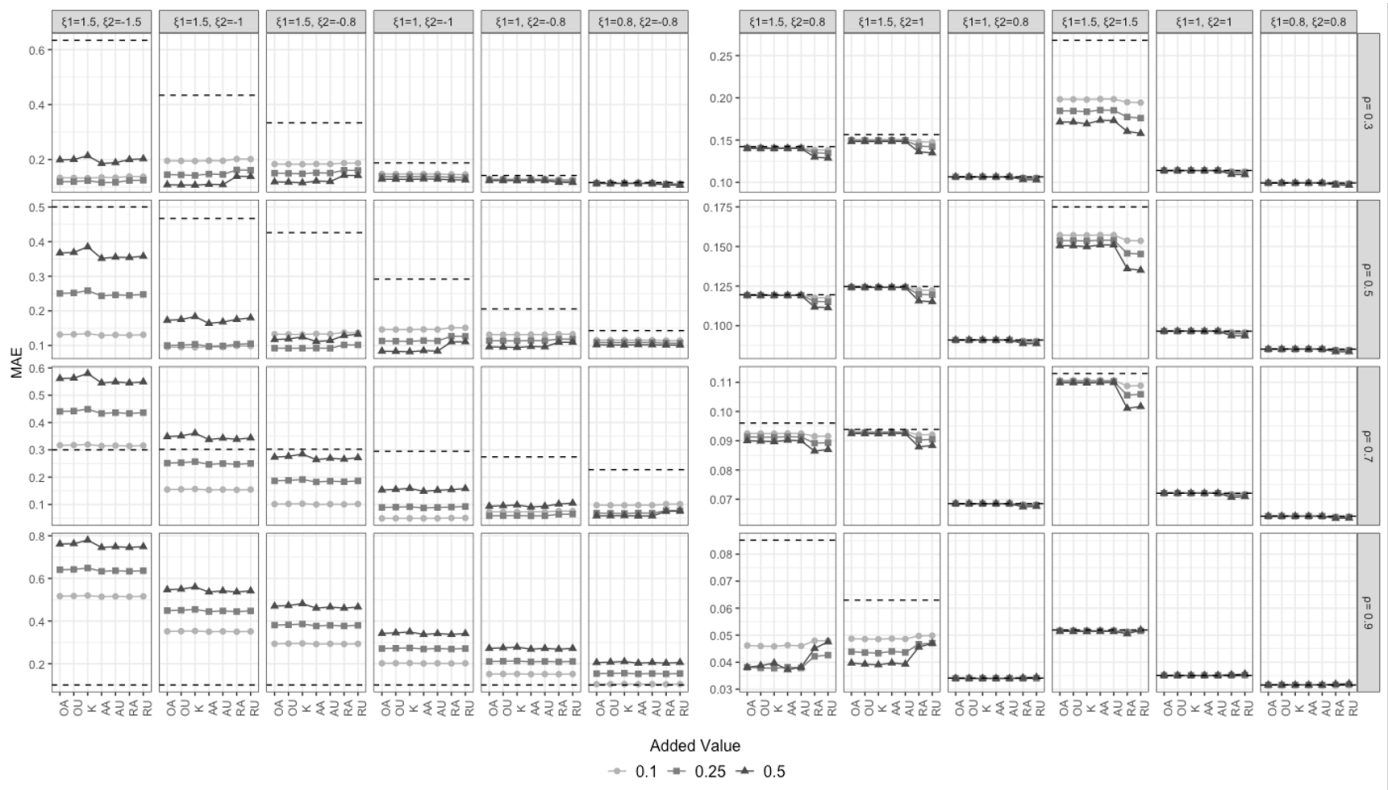


Figure A5: Mean Absolute Error (MAE) for Point Estimates (N=200)

Note. The structure of this figure is the same as Figure 2 in the paper.

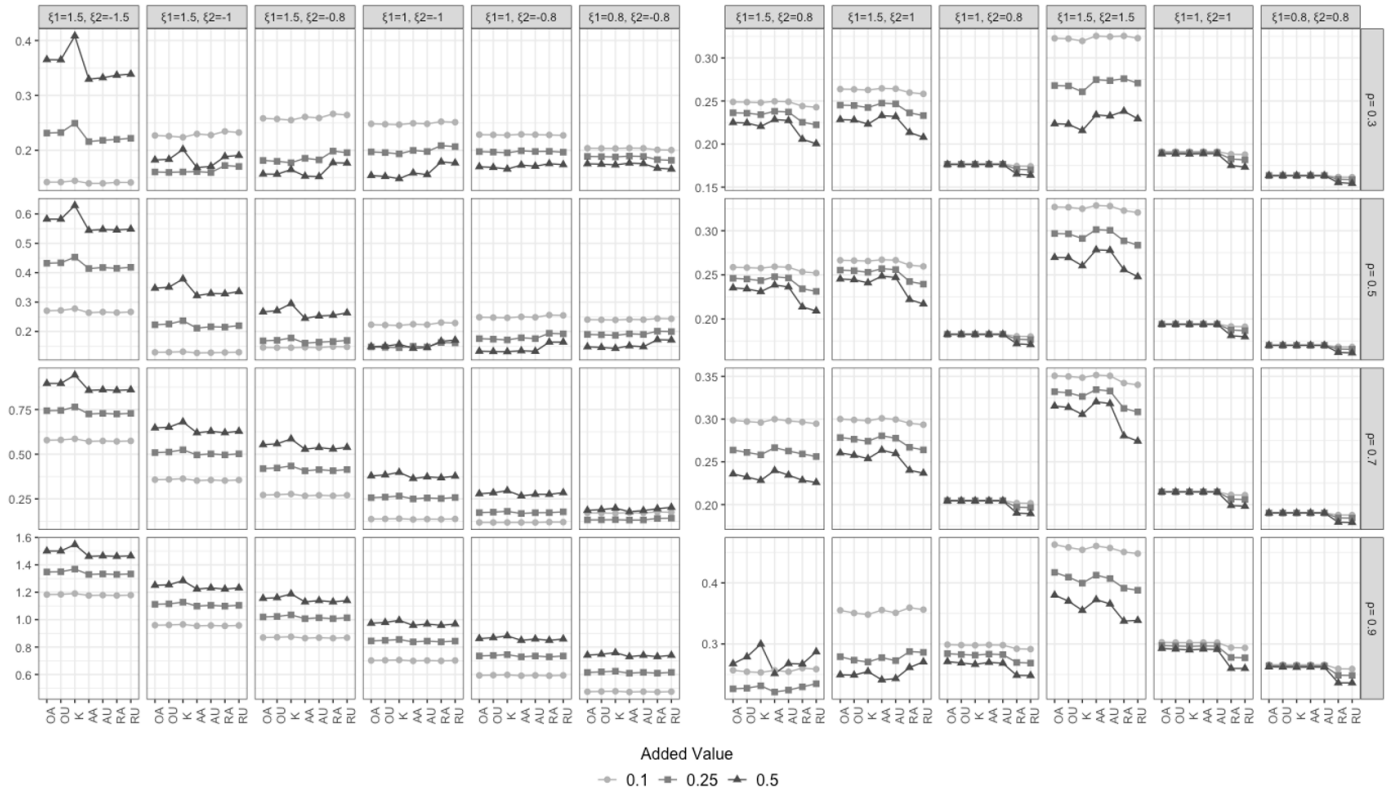


Figure A6: Mean Absolute Error (MAE) for Point Estimates after Fisher's Z-Transformation (N=100)

Note. The structure of this figure is the same as Figure 3 in the paper.

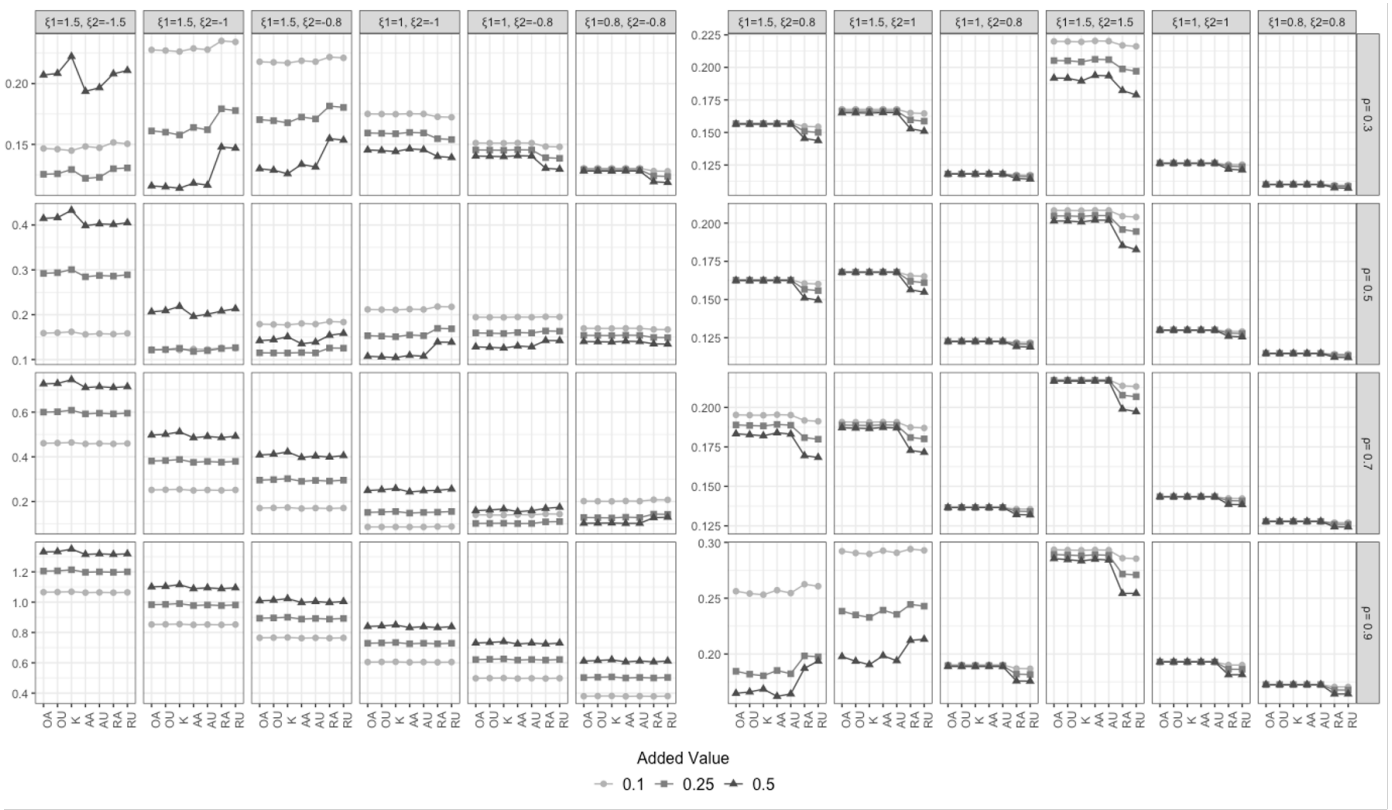


Figure A7: Mean Absolute Error (MAE) for Point Estimates after Fisher's Z-Transformation (N=200)

Note. The structure of this figure is the same as Figure 3 in the paper.

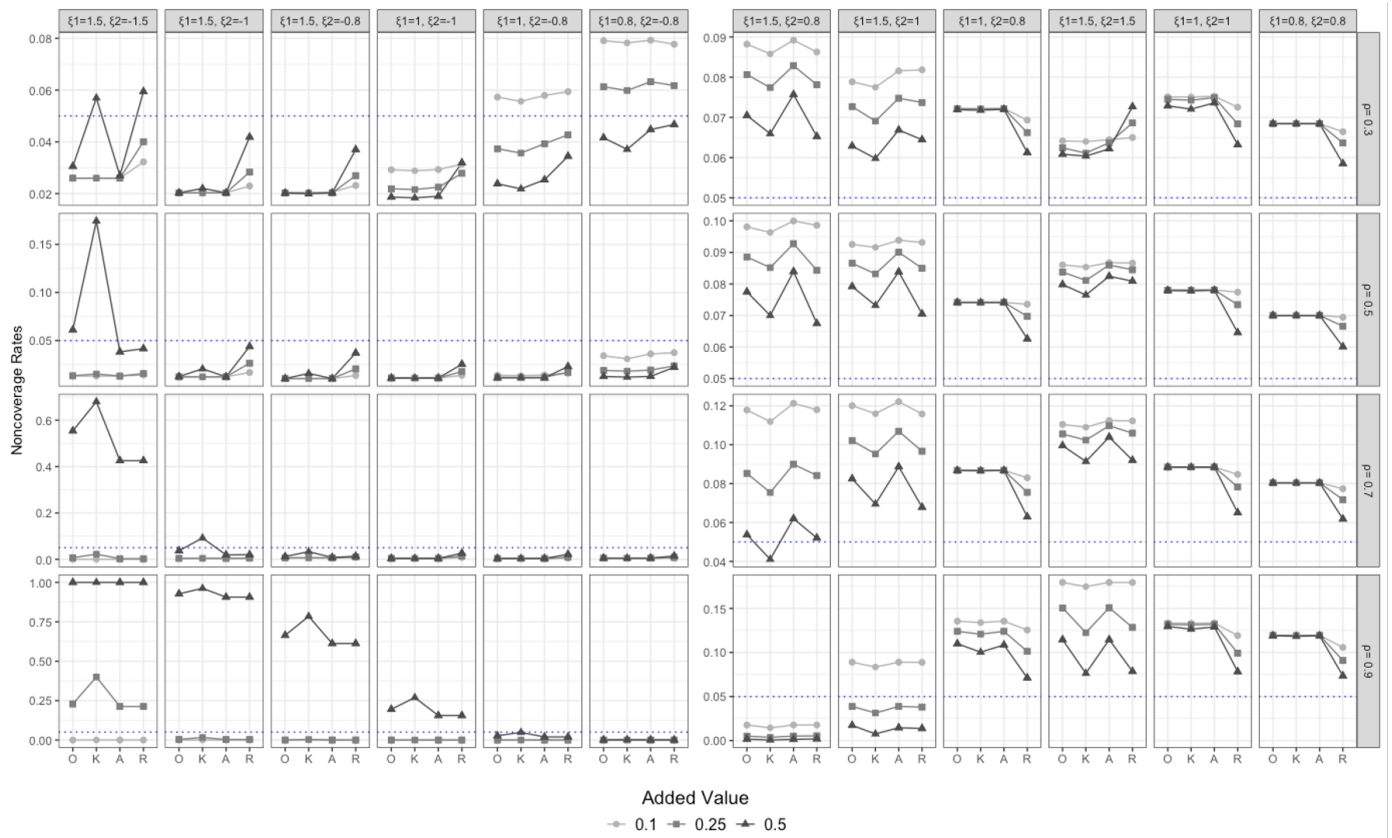


Figure A8: Noncoverage Rate of the 95% Wald Confidence Interval of the Correlation (N=100)

Note. The structure of this figure is the same as Figure 4 in the paper.

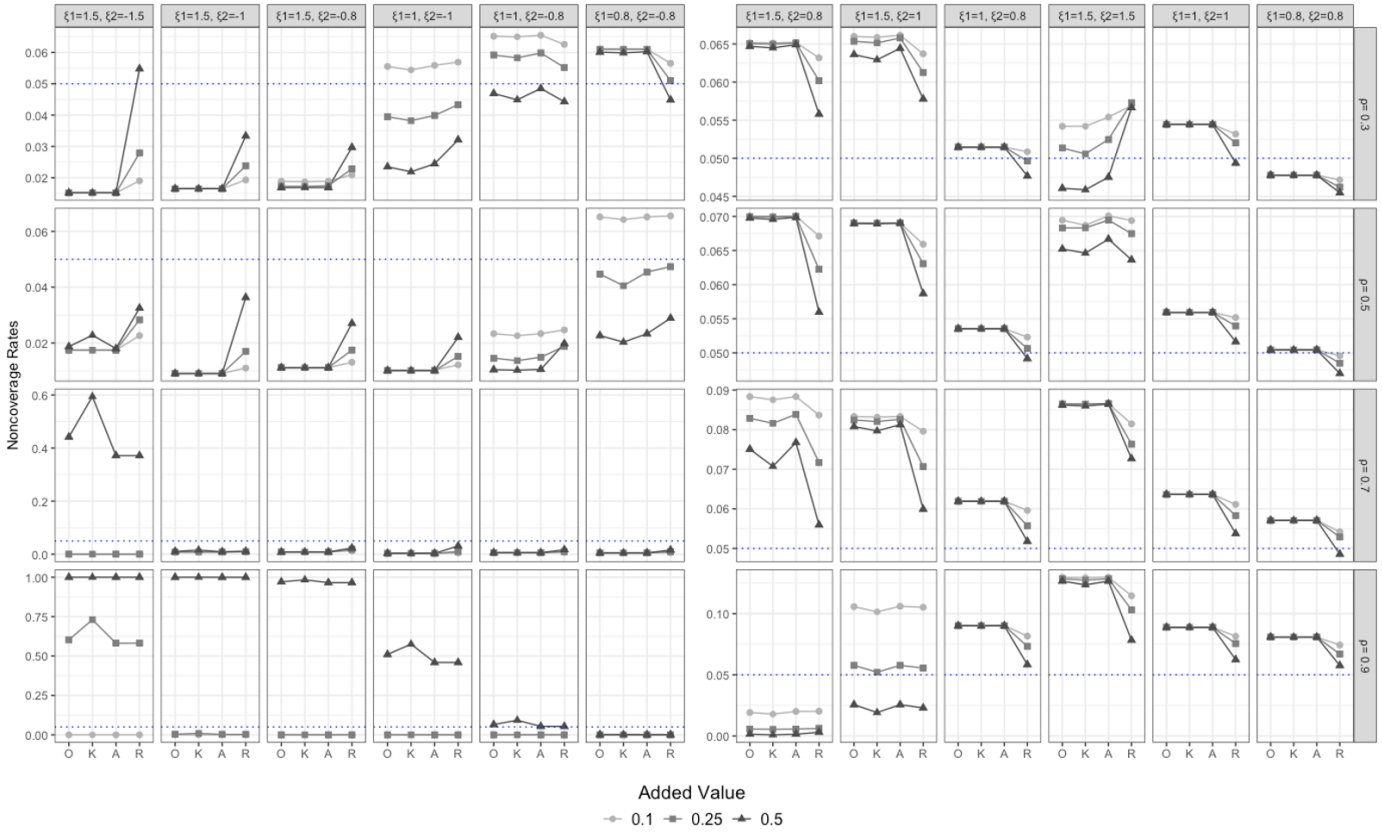


Figure A9: Noncoverage Rate of the 95% Wald Confidence Interval of the Correlation (N=200)

Note. The structure of this figure is the same as Figure 4 in the paper.