DOT/FAA/AM-00/15

Office of Aviation Medicine
Washington, DC 20591

Guidelines for Bootstrapping Validity Coefficients in ATCS Selection Research

Craig J. Russell
Michelle Dean
University of Oklahoma
Norman, Oklahoma 73019

Dana Broach
Civil Aeromedical Institute
Oklahoma City, Oklahoma 73125

May 2000

Final Report

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

U.S. Department of Transportation

Federal Aviation Administration

N O T I C E

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-00/15		

4. Title and Subtitle	5. Report Date
Guidelines for Bootstrapping Validity Coefficients in ATCS Selection Research	May 2000
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Russell, C.J., Dean, M.[1], and Broach, D.[2]	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
[1]University of Oklahoma, Norman, OK 73104 [2]Civil Aeromedical Institute , Oklahoma City, OK 73125	11. Contract or Grant No.

12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code

15. Supplemental Notes

This work was performed under task #AM-97-B-HRR-509.

16. Abstract

This technical report 1) reviews the literature on bootstrapping estimation procedures and potential applications to the selection of air traffic control specialists (ATCSs), 2) describes an empirical demonstration of procedures for estimating the sample size required to demonstrate criterion-related validity in ATCS selection, and 3) provides summary guidelines and recommendations for estimating sample size requirements in ATCS selection test validation using bootstropping procedures under conditions of direct and indirect range restriction. Bootstrapping estimates the sampling distribution of a statistic by iteratively resampling cases from a set of observed data. Confidence intervals are constructed for the statistic, providing an empirical basis for inferential statements about the likley magnitude of the statistic. Correlations between scores on the written ATCS aptitude test battery and subsequent performance in initial qualification training for a large sample of 10,869 controllers hired between 1986 and 1992 were bootstrapped in an empirical demonstration of the methodology. Finally, a three-step sequence of procedures is described for use in future bootstrap estimates of confidence intervals. Recommendations for sample size requirements in future ATC criterion validity studies include:

1. Results suggest samples of at least $N = 175$ to ensure the 90% confidence interval for r_{xy} does not contain 0.
2. Assumptions of bivariate normality in traditional parametric estimation procedures are not justified in the current data. Note that this observation may result in confidence intervals that are wider or narrower for any given sample size than intervals obtained from traditional parametric estimation.
3. Corrections for direct range restriction did not substantively influence whether the bootstrapped 90% confidence interval contained 0. Future applications should assess whether this holds true.
4. Given the apparent absence of bivariate normality in the current data, similar bootstrapping procedures should be used to assess whether the 90% confidence intervals for $\rho - \rho_o$ and $R_{Y.X1X2} - R_{Y.X1}$ contain 0.

Overall, the results suggest that bootstrapping of validity coefficients in controller selection research may be technically feasible. However, legal considerations may limit practical use of the methodology until accepted professional guidelines, standards, and principles are revised to accommodate innovative methodologies.

17. Key Words	18. Distribution Statement
Bootstrapping; Validity; Air Traffic Control Specialist Selection	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	38	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

GLOSSARY OF STATISTICAL SYMBOLS

B = Number of bootstrap iterations

H_o = Null hypothesis

n = Bootstrap sample size

N = Population size

r_b = Correlation computed on bootstrap sample

r_{xy} = Sample estimate of ρ

r_c = Estimate of ρ corrected for restriction in range on X

r'_{xy} = Estimate of ρ derived from the range - restricted sample

$R_{Y.X1}$ = Estimate of regression of X_1 on Y

$R_{Y.X1X2}$ = Estimate of regression of X_1 and X_2 on Y

ρ = Population or "true" Pearson product - moment correlation

SE_r = Standard error of the correlation

σ = Standard deviation of a score in sample

s_r^2 = Asymptotic variance of ρ

s_x = Standard deviation of X in the population

s'_x = Standard deviation of X in the range - restricted sample

$t_{\alpha/2}$ = Critical value of t in two - tailed test at the desired confidence level

X = Predictor score

$\overline{X}$ = Mean predictor score

Y = Criterion score

$\overline{Y}$ = Mean criterion score

GUIDELINES FOR BOOTSTRAPPING VALIDITY COEFFICIENTS IN ATCS SELECTION RESEARCH

INTRODUCTION

The Air Traffic Control Specialist (ATCS) occupation is the single largest (about 17,000 persons) and most publicly visible occupational group in the Federal Aviation Administration (FAA). Air traffic controllers are at the heart of a web of radars, computers, and communication facilities that comprise an increasingly complex and busy air transportation system. Competitive examinations have been used to determine entry into the occupation since 1964 (Brokaw, 1984). Validation of these competitive examinations has traditionally relied on concurrent, criterion-related designs with substantial samples of incumbent controllers. For example, about 800 incumbent controllers were sampled from 15 major cities in an early 1972 validation study. A subsequent longitudinal study drew data from over 2,300 controllers (Sells, 1984). The written test battery used between 1981 and 1992 for the selection of controllers was validated on samples ranging in size from 900 to over 3,000 (Boone, 1979). More recently, samples of 438 controller trainees and 296 incumbent controllers were used in predictive and concurrent criterion-related validation studies of a new generation of computer-administered tests for the occupation (Broach & Brecht-Clark, 1993).

Use of such large samples in validation studies imposes significant operational and financial burdens on the agency. For example, rearrangement of work schedules in field facilities is often required to allow controllers to participate in the studies and to ensure appropriate coverage of control positions. Consequently, overtime costs may be incurred by the facility to ensure adequate staffing during data collection efforts. Other incurred costs include (1) direct travel costs to bring the controller to the test site or the test to the controller, and (2) salary costs for the participating controllers. More efficient designs that require fewer controllers for selection test validation research are needed by the FAA to reduce the resource costs associated with validation of controller selection tests.

One possible approach is to maximize the information gained from a single sample of controllers using innovative, emerging statistical techniques, such as bootstrapping, to estimate the population validity coefficient for new selection tests. The fundamental task in criterion-related selection test validation is to make a probability-based inference about the magnitude of the "true" population validity coefficient, ρ, for a predictor, on the basis of a sample statistic, r_{xy}, computed on a sample of applicants or incumbents. Bootstrapping estimates the sampling distribution of a statistic by iteratively resampling cases, with replacement, from a set of observed data, and computing the sample statistic. Confidence intervals about the sample statistic can then be constructed, providing an empirical basis for inferential statements about the likely magnitude of the statistic. This approach allows for use of smaller samples for estimation of the underlying population parameter. Application of bootstrapping to estimation of validity coefficients might allow the FAA to user smaller samples in validation studies, thereby reducing resource costs.

Bootstrapping has been applied to the estimation of validity coefficients in methodological studies (Cooil, Winer, & Rados, 1987; Kromery & Hines, 1995). Other methodological studies have investigated the relationship of restriction in range and sample size to the accuracy of the confidence interval about a bootstrapped statistic (Allen & Dunbar, 1990; Mendoza, Hart, & Powell, 1991). However, bootstrapped estimates of criterion-related validity coefficients have not appeared in applied studies of personnel selection tests. Moreover, no guidelines or tables relating effect size (e.g., the magnitude of the r_{xy}), inferential errors (e.g., Type I and II errors), statistical power, and sample size have appeared for use by practitioners in designing selection test criterion-related validation studies with bootstrapping in mind. The purpose of this study was to develop

empirically-based guidelines and recommendations for estimating sample sizes required to attain reasonable and stable bootstrapped estimates of validity coefficients in concurrent, criterion-related validation of ATCS aptitude tests under conditions of explicit and incidental restriction in range.

TECHNICAL BACKGROUND

Traditional parametric estimation of sample size requirements

Correlation coefficient. The Pearson product moment correlation coefficient, ρ, reflects the strength of the linear relationship between two variables (Galton, 1888). A sample estimate of ρ is derived using the following formula:

$$r_{xy} = \frac{\sum (X - \overline{X}) \sum (Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} \quad \textbf{(Equation 1)}$$

Where

r_{xy} = sample estimate of ϱ

X = predictor score

$\overline{X}$ = mean predictor score

Y = criterion score

$\overline{Y}$ = mean criterion score

Brogden (1949) demonstrated 50 years ago that the economic utility of a personnel selection system is a direct function of the strength of the predictor-criterion (*X*-*Y*) relationship. More recently, Russell, Colella, and Bobko (1993) found that, if r_{xy} increases by a factor of 2, gross value-added to the organization doubles. Hence, an accurate sample estimate r_{xy} of the population parameter ρ provides key insight into how well a personnel selection system is working and its utility to the organization.

However, this begs the question, "How large does the sample have to be to ensure r_{xy} is an 'accurate' estimate of ρ?" Traditional means of answering this question use parametric assumptions about distributional characteristics of *X* and *Y*. For example, Fisher (1915, 1970, p. 194) described the asymptotic variance of a correlation (s_r^2) between two bivariate normally distributed variables *X* and *Y* as:

$$s_r^2 \cong \frac{(1 - \rho^2)^2}{N} \quad \textbf{(Equation 2)}$$

Where

N = sample size.

A minor variation of this formula yields the estimate of standard error (SE_r) used in the denominator of *t*-tests of H_o: $r_{xy} = 0$, or:

$$SE_r = \sqrt{\frac{(1 - r_{xy}^2)}{N - 2}} \quad \textbf{(Equation 3)}$$

The estimate for standard error of r_{xy} permits derivation of confidence intervals. The confidence interval (CI) is defined as

$$\text{CI} = r_{xy} \pm t_{a/2}(SE_r) \quad \textbf{(Equation 4)}$$

Where

$t_{\alpha/2}$ = the critical value of *t* in a two-tailed test at the desired confidence level (α = .10, *t* = 1.645 for the 90% confidence interval, or α = .05, *t* = 1.96 for the 95% confidence interval).

For example, the SE_r would be .07 for a sample of N = 200 and an ρ_{xy} of .20 between two bivariate normally distributed variables. One could be 90% sure the true population parameter ρ will fall in the interval .20 ± 1.645(.07), or .08 to .32. Similarly, one would be 95% sure the true population ρ will fall in the interval .20 ± 1.96(0.20), or .06 to .34. In this example, the 90% and 95% confidence intervals do not include zero, and one could reasonably infer the population correlation coefficient was not zero.

Assume the organization knows in advance that minimally acceptable criterion-related validity must be r_{xy} = .20 for a new personnel selection test to add economic value. One could then work backwards to determine the minimum sample size needed to ensure zero does not fall in the confidence interval (e.g., the null hypothesis H_o: $\rho_{xy} = 0$ can be rejected at p(type I error) ≤ .10) if in fact ρ = .20. For example, the minimum sample size needed to detect r_{xy} = .20 between two bivariate normal variables at p < .10 (2-tailed) can be obtained by solving for *N* as follows:

$$.20 = 1.645\sqrt{\frac{(1 - .20^2)}{N - 2}}, \text{ or, } N = 67 \text{ for a} = .10$$

Note, the median sample size of criterion validity studies reported in the *Journal of Applied Psychology* and *Personnel Psychology* between 1965 and 1991 was N = 104 (Russell et al., 1994).

Restriction in range. However, the distribution of predictor scores is generally non-normal in concurrent, criterion-related validity studies due to range restriction, as the predictor has been used to select the incumbents. For example, applicants to the ATCS occupation competed under civil service rules on the basis of a composite of written aptitude test scores (Broach, 1998). The distribution of that composite score for 205,592 ATCS applicants (out of over 400,000 since 1981) is presented in Figure 1. The distribution of that composite score for the 10,869 applicants competitively selected into the FAA between 1986 and 1992 is illustrated in Figure 2. Both figures superimpose what a normal curve with the same mean and standard deviation as data contained in the graph would look like. Note both distributions are distinctly non-normal and negatively skewed; the distribution of applicant composite scores (Figure 1) evidences some degree of bi-modality. The correlation between the composite score and subsequent performance in FAA Academy initial ATCS training for the 10,869 competitive entrants was r_{xy} = .182. However, the "true" population validity (ρ) is likely to be much larger than .182 in the N = 205,592 applicant population.

Ghiselli (1964) derived a correction formula that yields a more accurate estimate of ρ, i.e., what would have been expected if predictor and criterion data had been available on all applicants (Bobko & Rieck, 1980; Linn, Harnisch, & Dunbar, 1981). The formula correcting r_{xy} for direct range restriction is:

$$r_c = \frac{r'_{xy}\left(\frac{s_x}{s'_x}\right)}{\sqrt{1 - r'^2_{xy} + r'^2_{xy}\left(\frac{s_x}{s'_x}\right)^2}} \qquad \textbf{(Equation 5)}$$

Where

r_c = the estimate of ρ corrected for range restriction on X

r'_{xy} = the estimate of ρ derived from the range restricted sample

s'_x = the standard deviation of X in the range restricted sample

s_x = the standard deviation of X in the non - range restricted population

Application of this formula to the correlation between ATCS aptitude composite score and FAA Academy performance of r_{xy} = .182 yields:

$$r_c = \frac{.182\left(\frac{14.11}{5.02}\right)}{\sqrt{1 - .182^2 + .182^2\left(\frac{14.11}{5.02}\right)^2}} = \frac{.512}{1.23} = .42$$

Assumptions about the underlying distributions of *X* and *Y*. While direct range restriction on the predictor X constitutes a known violation of bivariate normality that can be "corrected" for, the X and Y distributions may be non-normal for any one of a large number of other reasons. Highly skewed (e.g., Figure 1) or multi-modal distributions of the predictor or criterion cause Fisher's bivariate normality assumption to be violated. For example, Fisher's formula assumes the sample was drawn from a single population characterized by a single value of ρ. Unfortunately, if the independent and dependent variables are distinctly non-normal, as appears to be the case in ATCS aptitude test scores, "tests . . . based on the large sample formula are often very deceptive" (Fisher, 1970, p. 195).

Moreover, it is possible that ATCS applicants were drawn from multiple populations, each with its own unique value of ρ. For example, Russell and Dean (1997) reported evidence of multiple population values of ρ in a sample of N > 15,000 applicants hired using the General Aptitude Test Battery (GATB) over a 10-year period. ATCS applicants were recruited from diverse demographic and geographic groups; it is possible that unique values of ρ might characterize the population of applicants from rural areas with just high school diplomas compared to the sub-population of applicants from large cities with college degrees. If applicants were drawn from multiple populations with unique ρ, using Fisher's formula to estimate confidence intervals and the sample sizes required to attain them will be incorrect.

In sum, correction formulae can be derived when deviations from bivariate normality are well understood, e.g., in the case of direct range restriction in the predictor. Unfortunately, when the distributional characteristics of X and Y are unknown, the underlying distributional characteristics of r_{xy} are also unknown, **and** Equation 2 cannot be used to estimate required sample size.

Bootstrap Estimation Procedures

Recently, Efron (1979) presented a new method of empirically estimating characteristics of population distributions from sample data, called bootstrapping. Bootstrapping estimates the sampling distribution of a statistic by iteratively resampling cases from a set of observed data. Basically, B "bootstrap" samples of size n are taken **with replacement** from the original sample of size N and saved to a file. An investigation using B = 1,000 bootstrap samples of size n will essentially be able to approximate the actual sampling distribution that would have been obtained if multiple independent samples of size N were drawn from the population. Bootstrapping is computationally time intensive, as the sample at hand is resampled with replacement many times to derive the distribution of the statistic of interest.

There are many advantages to using the bootstrap technique. First, it is not restricted to the normality assumptions of parametric tests. The percentile bootstrapping method (Efron & Tibshirani, 1993, chapter 13) generates confidence intervals directly from the bootstrapped sampling distribution (e.g., if B = 1,000 bootstrap samples are taken, the bootstrap correlations (r_b) representing the 5th and 95th percentile points would fashion the lower and upper points of a 90% CI). Of interest in this application is graphical interpretation of r_b frequency distributions (Efron & Tibshirani, 1993). Evidence of multimodality would suggest the presence of multiple subpopulations in the sample, each with a unique ρ. Second, information concerning the form of the original sample is retained, with no loss of distributional information. Rasmussen (1987) noted such loss of information does occur when nonparametric techniques convert data to ranks, which is why Lunneborg (1985) described bootstrapping as falling between parametric and nonparametric procedures for making probabilistic inferences.

The main disadvantage of the technique is that one must be confident the sample examined is indeed representative of the population from which it was drawn. Other than differences due to direct range restriction (which can be corrected for), this assumption appears to be met for studies of controller selection due to the large sample sizes. Regardless, applications of parametric statistical estimation procedures effectively make the same assumption. For example, if the test statistic of interest falls in the critical region, the investigator rejects the null hypothesis and proceeds to draw implications for theory and practice **as if** what is true in the sample is also true in the population. All inferential statistics must make the basic assumption that evidence drawn from the sample (e.g., rejection or lack of rejection of H_o) generalizes to the population.

An example. Rasmussen (1987) presented the following simple example to explain the bootstrap procedure. The computer initially is presented with a data set containing 10 graduate students' first year grade point average (GPA) and Graduate Record Exam (GRE) scores. Then a bootstrap sample (B_1) is randomly drawn with replacement from these 10 observations, causing the possibility of some observations being represented more than once in the bootstrap sample while other observations are not included. A single bootstrap sample may include the following cases: 5, 2, 8, 6, 2, 7, 9, 6, 1, and 2, resulting in a correlation of r_{b1} = .59. This procedure is repeated a large number of times (e.g., B = 1,000) and each r_b is saved to a separate file. The bootstrap correlations (r_b) are then rank ordered with the 50th and 950th correlations representing 90% confidence interval end points. The null hypothesis of $\rho_{GPA,GRE}$ = 0 is tested by determining whether 0 falls within the confidence interval (Rasmussen, 1987). The mean value of r_b across all B = 1000 bootstrap samples would be the best estimate of $\rho_{GPA,GRE}$.

Issues in bootstrapping. To determine the bootstrap's appropriateness, studies have been conducted examining the similarity in results between the bootstrap and traditional statistical approaches under conditions in which the parametric assumptions were met (e.g., Diaconis & Efron, 1983; Efron, 1985, 1986; Lunneborg, 1985). These studies resulted in bootstrap statistics (e.g., estimates of confidence intervals) that were extremely close to those

generated from traditional parametric approaches. Bickel and Freedman (1981; Freedman, 1981) demonstrated the bootstrap was asymptotically valid for many statistics (e.g., *t* and regression statistics).

A number of issues remain unresolved in using bootstrapping to conduct hypothesis testing. Most of these issues revolve around the relative accuracy of parametric versus bootstrap procedures in estimating probability intervals at the extreme tails of known (i.e., normal) distributions. However, the percentile method of estimating confidence intervals, as described by Efron and Tibshirani (1993), provides "good theoretical coverage properties as well as reasonable stability in practice" (p. 169). Good "theoretical coverage" refers to confidence intervals that 1) accurately estimate the probability of the population parameter falling within the confidence interval **and** 2) divide "coverage error" equally across the two tails.

Hence, the percentile bootstrap method of estimating confidence intervals might be used to estimate the distributional characteristics of r_{xy} under actual conditions faced by the FAA in the selection of air traffic controllers. In this study, bootstrap procedures were applied to archival ATCS selection data to estimate sampling distributions for r_{xy} obtained with $B = 1{,}000$ samples of $n = 25, 50, 75, \ldots$ to 200. Results are presented with and without correction for direct range restriction.

METHOD

Sample

The Civil Aeromedical Institute provided archival ATCS written aptitude test scores for 205,592 examinations for the period 1981 to 1992. The Institute also provided test and criterion data for the 10,869 persons competitively selected into the ATCS occupation from October 1985 through January 1992.

Measures

Predictor. The written ATCS aptitude test battery consisted of three tests: (a) the Multiplex Controller Aptitude Test (MCAT); (b) the Abstract Reasoning Test (ABSR); and (c) the Occupational Knowledge Test (OKT). The MCAT was a timed, 110-item paper-and-pencil civil service test (OPM test No. 510) simulating activities required for control of air traffic. Multiple, parallel forms of these test were available (Lilienthal & Pettyjohn, 1981). Aircraft locations and direction of flight were indicated with graphic symbols on a simplified, simulated radar display (Figure 3). An accompanying table provided relevant information required to answer the item, including aircraft altitudes, speeds, and planned routes of flight. MCAT test items required examinees to identify situations resulting in conflicts between aircraft, interpret tabular and graphical information , and to solve time, speed, and distance problems. The ABSR was a timed, multiple-choice, 50-item civil service examination (OPM test No. 157). To solve an item, examinees determined what relationships existed within sets of symbols or letters. The examinee then identified the next symbol or letter in the progression, or the element missing from the set. A sample ABSR item is presented in Figure 4. The OKT was a timed, multiple-choice 80-item job knowledge test that contained items related to seven knowledge domains relevant to aviation, generally, and to air traffic control phraseology and procedures, specifically. The OKT was developed as an alternative to self-reports of aviation and air traffic control experience. The OKT was found to be more predictive of performance in ATCS training than self-reports (Dailey & Pickrel, 1984; Lewis, 1978).

The development of the written ATCS aptitude test battery has been extensively described elsewhere (Brokaw, 1984; Collins, Boone, & VanDeventer, 1984; Manning, 1991; Sells, 1984; Sells, Dailey, & Pickrel, 1984). The test-retest correlation for the MCAT was estimated at .60 in a sample of 617 newly hired controllers (Rock, Dailey, Ozur, Boone, & Pickrel, 1982, p. 59). Parallel form reliability, as computed on the same sample, ranged from .42 to .89 for various combinations of items (Rock et al., p. 103). Lilienthal and Pettyjohn (1981) examined internal consistency and item difficulties for 10 versions of the MCAT. Cronbach's alpha ranged from .63 to .93; the alphas for 7 of the 10 versions were greater than .80. In contrast, no item analyses, parallel form, test-retest, or internal consistency estimates of the ABSR test have been reported.

Weighted MCAT and ABSR raw scores were summed and transformed to a score with a mean of 70 and maximum of 100, known as the Transmuted Composite Score (TMC). No estimates for the reliability of this composite score have been reported. About half of all applicants were expected to score at or above the mean (Rock, Dailey, Ozur, Boone, & Pickrel, 1984).

Criterion. The criterion was performance in the FAA Academy initial ATCS training program, known as the ATCS Non-radar Screen ("the Screen"). Under the *Uniform Guidelines for Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978), training may be used as a criterion measure where success in training is "properly measured," and the relevance of the training can be demonstrated through comparison of training content to critical or important job behaviors, or by showing that training measures are related to subsequent measures of job performance. The Screen was originally established in response to recommendations made by the U.S. Congress House Committee on Government Operations (U.S. Congress, 1976) to "...provide early and continued screening to insure (sic) the prompt elimination of unsuccessful trainees and relieve the regional facilities of much of this burden" (p. 13). The Screen was based upon a miniaturized training-testing-evaluation personnel selection model (Siegel, 1978, 1983; Siegel & Bergman, 1975) in which individuals with no prior knowledge of an occupation are trained and then assessed for their potential to succeed in the job. Performance in the Screen has been shown to predict subsequent performance in radar based training one to two years after entry into the occupation (Broach & Manning, 1994), as well as completion of the rigorous on-the-job training sequence and certification as a qualified "full performance level" (FPL) controller (Broach, 1998; Della Rocco, 1998; Della Rocco, Manning, & Wing, 1990; Manning, Della Rocco, & Bryant, 1989).

Thirteen assessments of performance, including six classroom tests, observations of performance in six laboratory simulations of non-radar air traffic control, and a final written examination, were made during the Screen (Broach, Farmer, & Young, in review; Della Rocco, 1999; Della Rocco, Manning, & Wing, 1990). The final summed composite score (NLCOMP) was weighted 20% for the classroom tests, 60% for laboratory simulations, and 20% for the final examination. A minimum NLCOMP score of 70 was required to pass. The final composite score was the criterion measure in this study.

Bootstrap procedures

The SYSTAT7.01 statistical package, published by SPSS Inc., was used for all derivations (syntax files are available from the first author). Each bootstrap procedure reported below followed a four-step sequence to yield "percentile" confidence intervals, as described by Efron and Tibshirani (1993).

Step1: Number of iterations. Decide how many bootstrap samples (B) to take. Evidence suggests bootstrap estimates of common statistics' distributional characteristics tend to stablize when the number of bootstrap samples drawn approaches $B = 200$ (Efron & Tibshirani, 1993, p. 52). However, point estimates of confidence interval percentiles (e.g., the 5th and 95th percentiles) are subject to greater error in estimation. Efron and Tibshirani recommend extracting 500 to 1,000 bootstrap samples to minimize estimation error (1993, p. 252). Hence, to ensure accuracy, all bootstrap procedures reported here iteratively drew $B = 1{,}000$ bootstrap samples with replacement.

Step 2: Number of sampled observations for bootstrap. Decide how many observations should be drawn in each of the B_1 to B_{1000} bootstrap samples. Given the parametric estimate of sample size required in the current data for the sample stastistic $r_{xy} = .182$ to reject H_o: $\rho = 0$ at $\alpha = .10$ was $N = 81$ (as derived from Equation 3), eight independent bootstraps of $n = 25$ through $n = 200$ were performed. In other words, first, $B = 1{,}000$ samples of size $n = 25$ were drawn, with replacement, from the original sample of $N = 10{,}869$. Then $B = 1{,}000$ samples of size $n = 50$ were drawn, with replacement, from the original sample, followed by $B = 1{,}000$ samples of size $n = 75$, $B = 1{,}000$ samples of size $n = 100$, and so forth until a total of eight bootstrap operations had been performed for $n = 25, 50, 75, \ldots, 200$.

Step 3: Compute bootstrapped statistic. The TMC-NLCOMP Pearson product moment correlation (r_b) and TMC standard deviation were derived for each bootstrap sample (B_1 to B_{1000}) and saved to a file labeled TMC25. This procedure was repeated independently for $n = 50, 75, 100, 125, 150, 175$, and 200, yielding additional output files labeled TMC50 through TMC200.

Step 4: Examine distribution of bootstrapped statistic. Correlations (r_b) derived from each bootstrap procedure were sorted and values corresponding to the 5th, 50th, and 95th percentile identified. The frequency with which each r_b value occured was then plotted graphically, with the 5th, 50th, and 95th percentile values labeled below the X-axis.

Analyses

Uncorrected correlations. Three analyses were performed to generate different distributions of r_b for each bootstrap sample size (n = 25, 50, … 200). A frequency distribution of r_b was plotted and the 5th, 50th, and 95th percentile values for the simple, uncorrected TMC-NLCOMP correlation were derived. For comparison purposes, the normal curve with a mean and standard deviation identical to that found in the r_b frequency distribution was superimposed. Basic sampling theory predicts that the interval between the 5th and 95th percentile values of r_b will decrease as sample size increases. The smallest bootstrap sample size (n) with a 90% confidence interval that no longer contains 0 will approximate the minimum sample size (N) needed to ensure r_{xy} = .182 will reject H_o: $\rho = 0$ at $\alpha = .10$. Computational time required for this procedure ranged from two hours (bootstrap n = 25) to 6 hours (bootstrap n = 200) on a 233 Mhz Intel Pentium® personal computer. Graph A-1 in Appendix A portrays the frequency distribution output for B = 1,000 bootstrap samples of size n = 25 for the simple, uncorrected TMC-NLCOMP correlation. Graphs A-2 through A-8 in Appendix A present the frequency distributions for r_b derived for B = 1,000 bootstrap samples of n = 50, 75, 100, 125, 150, 175, and 200, respectively. The logical flow of this analysis is illustrated in Figure 5.

Correlations corrected for restriction in range. Second, Ghiselli's (1964) correction formula for direct range restriction was applied to each r_b within the TMC25, TMC50, … and TMC200 files. The TMC standard deviation (s'_x) for each bootstrap sample was computed and saved to the file with each respective r_b. Subsequently, each r_b was corrected using s'_x derived from the bootstrap sample from which it was drawn, and s_x = 14.11, derived from the N = 206,592 applicant population. The corrected bootstrapped correlation coefficients (r_b) were rank ordered and plotted, yielding a frequency distribution with the 5th, 50th, and 95th percentile points indicated on the X-axis. The flow of this analysis is illustrated in Figure 6. Again, for purposes of comparison, a normal curve with a mean and standard deviation identical to that found in the corrected r_b frequency distribution was superimposed on the r_b frequency distribution. Graph B-1 in Appendix B is the result of this procedure applied to the B = 1,000 bootstrap samples of size n = 25. Graphs B-2 through B-8 in Appendix B present the frequency distributions for r_b corrected for restriction in range for B = 1,000 bootstrap samples of n = 50, 75, 100, 125, 150, 175, and 200, respectively.

Correlations generated for bivariate normal population. Finally, using the SYSTAT7.01 random normal function, 1,000 r_b were generated from B = 1,000 samples of n = 25 taken from a bivariate normal population with ρ = .182. The standard deviation was computed as σ = .1974, based on Equation 2. These bivariate normal bootstraped correlation coefficients were rank ordered and plotted, yielding a frequency distribution with the 5th and 95th percentile points indicated on the X-axis, as were the values ρ = .182 and σ = .1974 used to generate the data. Graph C-1 in Appendix C presents the result of this procedure. This procedure was repeated for samples of n = 50, 75, … 200, computing the standard deviation each time based on equation 2. The flow of this analysis is portrayed in Figure 7. Graphs C-2 through C-8 in Appendix C present the frequency distributions for r_b derived for B = 1,000 bootstrap samples of n = 50, 75, 100, 125, 150, 175, and 200, respectively.

In sum, the graphs in Appendix A capture the bootstrapped r_b frequency distribution for TMC-NLCOMP correlations uncorrected for range restriction. The Appendix B graphs capture the bootstrapped r_b frequency distribution for TMC-NLCOMP correlations corrected for range restriction. Last, the graphs in Appendix C are what a bootstrapped r_b frequency distribution for TMC-NLCOMP correlations is expected to look like if the applicant population was characterized by bivariate normal distribution of TMC and NLCOMP and $\rho_{TMC,\ NLCOMP}$ = .182. Confidence interval end points for corrected and uncorrected bootstrap procedures are summarized in Table 1.

RESULTS

A number of inferences can be drawn from the graphs and their respective confidence intervals. First, and perhaps most obvious, visual interpretation suggests that the distributional characteristics of r_b (corrected or uncorrected for direct range restriction) are not what would be expected under conditions of bivariate normality — the "C" graphs differ meaningfully from the "A" and "B" graphs. Hence, for the

90% confidence interval or a = .10, the estimated N = 81 sample size required to detect ρ = .182 derived under parametric assumptions is spurious.

Second, examination of the confidence intervals summarized in Table 1 indicates $N \approx 175$ or greater is required to ensure the 90% confidence interval does not contain 0 (i.e., H_o: $\rho = 0$ will be rejected) in these archival data. The actual distributional characteristics of these data, as revealed by the bootstrap procedure, suggest a larger sample ($N \geq 175$) will be required to reject H_o: $\rho = 0$ than would be required if the joint TMC-NLCOMP space was bivariate normal (N = 81).

Third, given the relatively tight range of observed TMC values in the N = 10,896 competitively selected controllers, virtually no outliers were present. Bootstrapping procedures are most subject to estimation error when the original sample contains infrequent, extreme outliers (Efron & Tibshirani, 1993); Figure 2 indicates this was not a problem in the current data.

Fourth, the same cannot be said of TMC values in the original applicant pool, which suggest a small group of extremely low TMC values lie some distance from the rest of the observations. These outliers will inflate the non-range restricted standard deviation estimate (s_x) in Ghiselli's (1964) correction formula. This may have been due to labor pool "history effects" associated with the Professional Air Traffic Controller Organization (PATCO) strike of the early 1980s. That is, there may have been a higher than usual frequency of low-ability, unsuccessful applicants attracted by the publicity about the ATCS occupation following the strike. If the outliers were due to such a history effect, Ghiselli's correction for range restriction may represent a spurious overcorrection when estimating ρ in future applicant pools that are not influenced by a similar history effect.

Finally, noting that this last caveat holds for all inferences drawn from the current analyses —these results will generalize to future criterion validation efforts only to the extent that similar TMC-NLCOMP distributional characteristics and latent TMC-NLCOMP relationships exist.

Guidelines and Recommendations

A number of recommendations can be drawn for future FAA efforts at estimating criterion validity. First, extremely large (1,000+) sample sizes are not required to yield accurate estimates of selection battery criterion validity. Results suggest samples in the range of N = 200-500 ought to provide whatever margin of error might be needed to ensure accurate estimation of ρ, i.e., to ensure 0 does not fall in the 90% confidence interval. A number of additional recommendations and guidelines follow:

1. Assumptions of bivariate normality in traditional parametric estimation procedures are not justified in the current data. Estimation of confidence intervals and tests of null hypotheses should be performed using the four-step bootstrap procedure outlined above. Note that this recommendation may result in confidence intervals that are larger or smaller than those obtained from traditional parametric estimation for any given sample size.
2. Corrections for range restriction did not substantively influence whether the bootstrap estimated 90% confidence interval contained 0. Future applications should continue to assess whether this holds true. Note, under parametric assumptions, the estimate of the standard deviation of r_c is:

$$SD(r_c) = \frac{\frac{s_x}{s'_x}}{\sqrt{\left[1 + r'^2\left(\frac{s_x^2}{s'^2_x} - 1\right)\right]^{3/2}}} SD(r')$$

Equation (6)

Where

s_x = the standard deviation of X in the unrestricted population

s'_x = the standard deviation of X in the range restricted sample

s_x^2 = the variance of X in the unrestricted population

s'^2_x = the variance of X in the range restricted sample

r'^2 = the squared observed correlation in the range restricted sample

r' = the observed correlation in the range restricted sample.

This correction can then be used, again under parametric assumptions, to test H_o and derive confidence intervals for r_c. Importantly, while $SD(r)_c$ is larger than $SD(r)$ under conditions of range restriction, $SD(r_c)$ does not increase relative to $SD(r)$ as fast as r_c increases relative to r (Bobko, 1995). Hence, under parametric conditions the

investigator enjoys a "boost" in statistical power when testing H_o: $r_c = 0$. Current results suggest this "boost" is not justified in the present data; the likelihood of 0 falling in the confidence interval seems to be about the same for both r and r_c.

3. Given the apparent absence of bivariate normality in the current data, tentative implications can also be drawn for tests of H_o: $\rho = \rho_o \neq 0$ and of H_o: $R_{Y.X1} \geq R_{Y.X1X2}$. Specifically, parametric tests of H_o: $\rho = \rho_o \neq 0$ require use of Fisher's Z transformation. In the presence of a constant effect size, the resultant Z test (Bobko, 1995, p. 54) literally requires **doubl e** the sample size to attain the same statistical power as a test of H_o: $\rho = 0$. Again, the absence of bivariate normality suggested by the current results implies similar bootstrapping procedures should be used to assess whether the 90% confidence intervals for $\rho - \rho_o$ and $R_{Y.X1X2} - R_{Y.X1}$ contain 0.

Overall, these results indicate that accurate estimation of validity coefficients by bootstrap may be technically feasible. However, two factors may limit the practical application of the method at present. First, current professional guidelines, standards, principles, and practices in selection test validation are based on traditional parametric statistics. Further methodological research and empirical demonstrations must be conducted to provide the technical foundation for revising these professional canons. Second, personnel selection tests and their validation are subject to legal review. Statistical evidence in employment discrimination litigation has probative value only to the degree that the underlying theory, model, and method are credible (Howard, 1994). Bootstrap is not without controversy, and therefore may not be viewed as credible in litigation. The principal challenge may come from the *Uniform Guidelines*' admonition to "… avoid reliance upon techniques which to tend to overestimate a validity finding as a result of capitalization on chance … ." If not carefully explained, the bootstrap may have the appearance of exploiting chance to an unprecedented degree. Further research is required to demonstrate that the bootstrap does not lead to overestimates of validity, nor does it capitalize on chance.

REFERENCES

Allen, N. L., & Dunbar, S. B. (1990). Standard errors of correlations adjusted for incidental selection. *Applied Psychological Measurement*, *14*, 83 - 94.

Bickel, P., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics, 9*, 1196-1217.

Bobko, P. (1995). *Correlation and regression: Principles and applications of industrial/organizational psychology and management*. New York: McGraw-Hill Inc.

Bobko, P. & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement, 4*, 385-98.

Boone, J. O. (1979). *Toward the development of a new selection battery for air traffic control specialists.* (DOT/FAA/AM-79/21). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA080065/6).

Broach, D. (1998). Air traffic control specialist aptitude testing, 1981-1992. In D. Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992* (pp. 7-16). (DOT/FAA/AM-98/23). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Broach, D., & Brecht-Clark, J. (1993). Validation of the Federal Aviation Administration air traffic control specialist Pre-Training Screen. *Air Traffic Control Quarterly*, *1*, 115 - 33.

Broach, D., Farmer, W. L., & Young, W. C. (In review). *Differential prediction of FAA Academy performance on the basis of race and written air traffic control specialist aptitude test scores.* Oklahoma City, OK: Federal Aviation Administration Civil Aeromedical Institute.

Broach, D., & Manning, C. A. (1994). *Validity of the air traffic control specialist Non-radar Screen as a predictor of performance in radar-based air traffic control training.* (DOT/FAA/AM-94/9). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA279745)

Brogden, H. (1949). When testing pays off. *Personnel Psychology, 2*, 171-83.

Brokaw, L. D. (1984). Early research on controller selection. In S. B. Sells, J. T. Dailey, & E. W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 39-78). (DOT/FAA/84-2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA147765).

Cooil, B, Winer, R.B., & Rados, D.L. (1987). Cross-validation for prediction. *Journal of Marketing Research, 24*, 271-9.

Collins, W. E., Boone, J. O., & VanDeventer, A. W. (1984). The selection of air traffic control specialists: Contributions by the Civil Aeromedical Institute. In S. B. Sells, J. T. Dailey, & E. W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 79 - 112). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. 147765).

Dailey, J. T., & Pickrel, E. W. (1984). Development of the Multiplex Controller Aptitude Test. In S. B. Sells, J. T. Dailey, & E. W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 281 - 297). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA147765).

Della Rocco, P. S. (1998). FAA Academy air traffic control specialist screening programs and strike recovery. In D. Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992* (pp. 17-22). (DOT/FAA/AM-98/23). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Della Rocco, P. S., Manning, C. A., & Wing, H. (1990). *Selection of controllers for automated systems: Applications from current research.* (DOT/FAA/AM-90/13). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA230058).

Diaconis, P. & Efron, B. (1983, May). Computer-intensive methods in statistics. *Scientific American*, 116-30.

Efron, B. (1979). Computers and the theory of statistics: Thinking the unthinkable. *Society for Industrial and Applied Mathematics Review, 21*, 460-80.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika, 72*, 45-58.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association, 81*, 461-70.

Efron, B. & Tibshirani R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures, 29 *CFR* 1607.

Fischer, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*, 507-21.

Fisher, R.A. (1970). *Statistical methods for research workers* (14th edition). Hafner Press: New York.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics, 9,* 1218-28.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, Vol. XLV, 135-45.

Ghiselli, E. (1964). *Theory of psychological measurement*. New York: McGraw-Hill.

Howard, W. M. (1994). The decline and fall of statistical evidence as proof of employment discrimination. *Labor Law Journal, 45*, 208-20.

Kromery, J.D. & Hines, C.V. (1995). Use of empirical estimates of shrinkage in multiple regression: A caution. *Education and Psychological Measurement, 55*, 901-25.

Lewis, M. A. (1978). *Use of the Occupational Knowledge Test to assign extra credit in selection of air traffic controllers.* (DOT/FAA/AM-78/7). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA05367/5GI).

Lilienthal, M. G., & Pettyjohn, F. S. (1981). *Multiplex Controller Aptitude Test and Occupational Knowledge Test: Selection tools for air traffic controllers.* (NAMRL Special Report 82-1). Pensacola, FL: Naval Aerospace Medical Research Laboratory. (NTIS No. ADA118803).

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology*, *66*, 655-63.

Lunneborg, C. E. (1985). Estimating the correlation coefficient: The bootstrap approach. *Psychological Bulletin, 98,* 209-15.

Manning, C. A. (1991). Procedures for selection of air traffic control specialists. In H. Wing & C. A. Manning (Eds.), *Selection of air traffic controllers: Complexity, requirements, and public interest* (pp. 13 - 22). (DOT/FAA/AM-91/9). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA238267).

Manning, C. A., Della Rocco, P. S., & Bryant, K. (1989). *Prediction of success in air traffic control field training as a function of selection and screening performance.* (DOT/FAA/AM-89/6). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. *Multivariate Behavioral Research*, *26*, 255-69.

Rasmussen, J.L. (1987). Estimating correlation coefficients: Bootstrap and parametric approaches. *Psychological Bulletin, 101,* 136-9.

Rock, D. B., Dailey, J. T., Ozur, H., Boone, J. O., & Pickrel, E. W. (1982). *Selection of applicants for the air traffic controller occupation.* (DOT/FAA/AM-82/11). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA122795/8).

Rock, D. B., Dailey, J. T., Ozur, H., Boone, J. O., & Pickrel, E. W. (1984). Research on the experimental test battery for ATC applicants: Study of job applicants, 1978. In S. B. Sells, J. T. Dailey, & E. W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 411-58). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. 147765).

Russell, C. J., Colella, A., & Bobko, P. (1993). Expanding the context of utility: The strategic impact of personnel selection. *Personnel Psychology, 46*, 781-801.

Russell, C. J., Settoon, R. P., McGrath, R., Blanton, A. E., Kidwell, R. E., Lohrke, F. T., Scifries, E.L., & Danforth, G.W. (1994). Investigator characteristics as moderators of selection research: A meta-analysis. *Journal of Applied Psychology, 79*, 163-70.

Russell, C. J. & Dean, M. A. (1997, April). *In search of situation specificity*. Presented at the 12th annual meetings of the Society of Industrial and Organizational Psychology, St. Louis, MO.

Sells, S. B. (1984). Validation of the new ATCS selection tests on trainees and controller populations: Three studies — 1972, 1977, 1979. In S. B. Sells, J. T. Dailey, & E. W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 353-395). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA14765).

Siegel, A. I. (1978). Miniature job training and evaluation as a selection/ classification device. *Human Factors*, *20*, 189-200.

Siegel, A. I. (1983). The miniature job training and evaluation approach: Additional findings. *Personnel Psychology*, *36*, 41-56.

Siegel, A. I., & Bergman, B. A. (1975). A job learning approach to performance prediction. *Personnel Psychology*, *28*, 325-39.

U. S. Congress. (January 20, 1976). *House Committee on Government Operations recommendations on air traffic control training.* Washington, DC: Author.

FIGURES

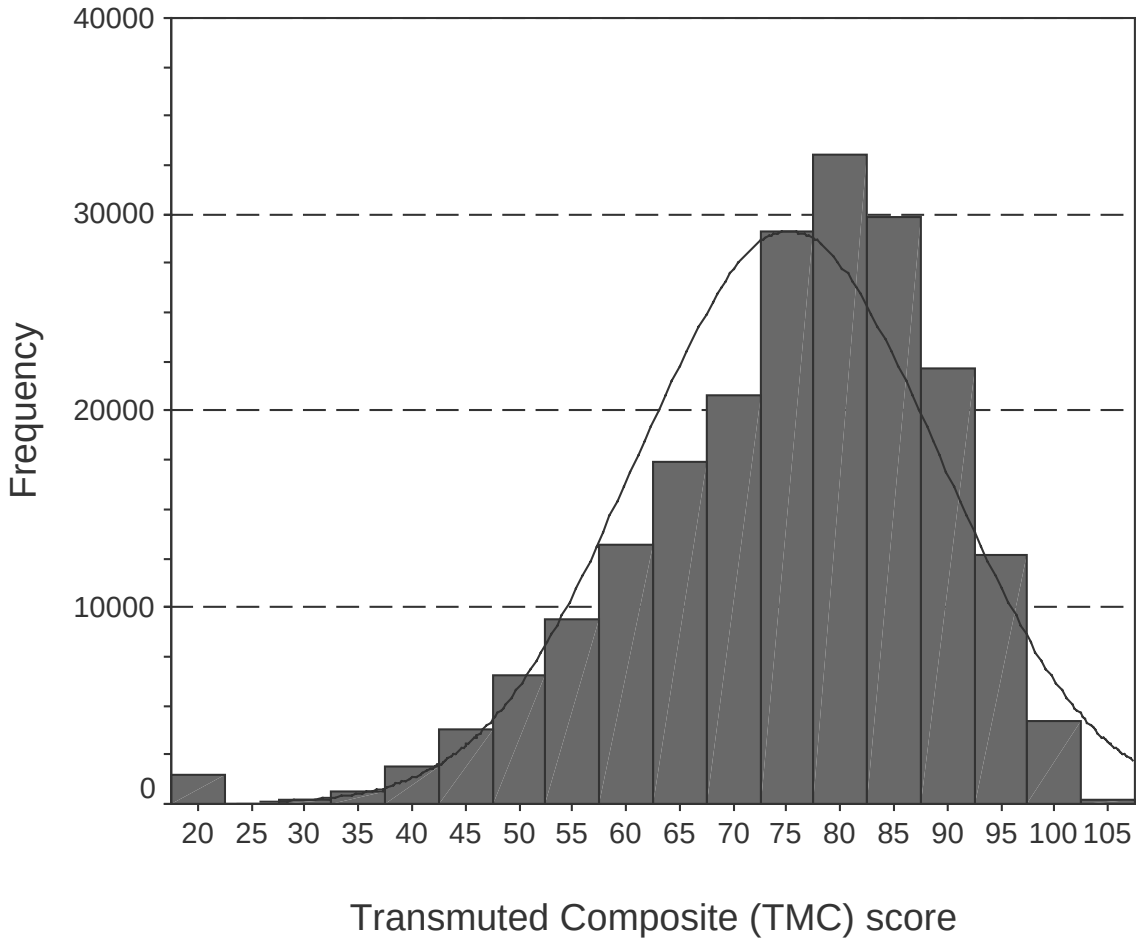


Figure 1: Frequency distribution of TMC scores in unrestricted (Applicant) sample *N* = 206,592 **(** $\overline{X} = 75.20, \sigma = 14.11$**)**

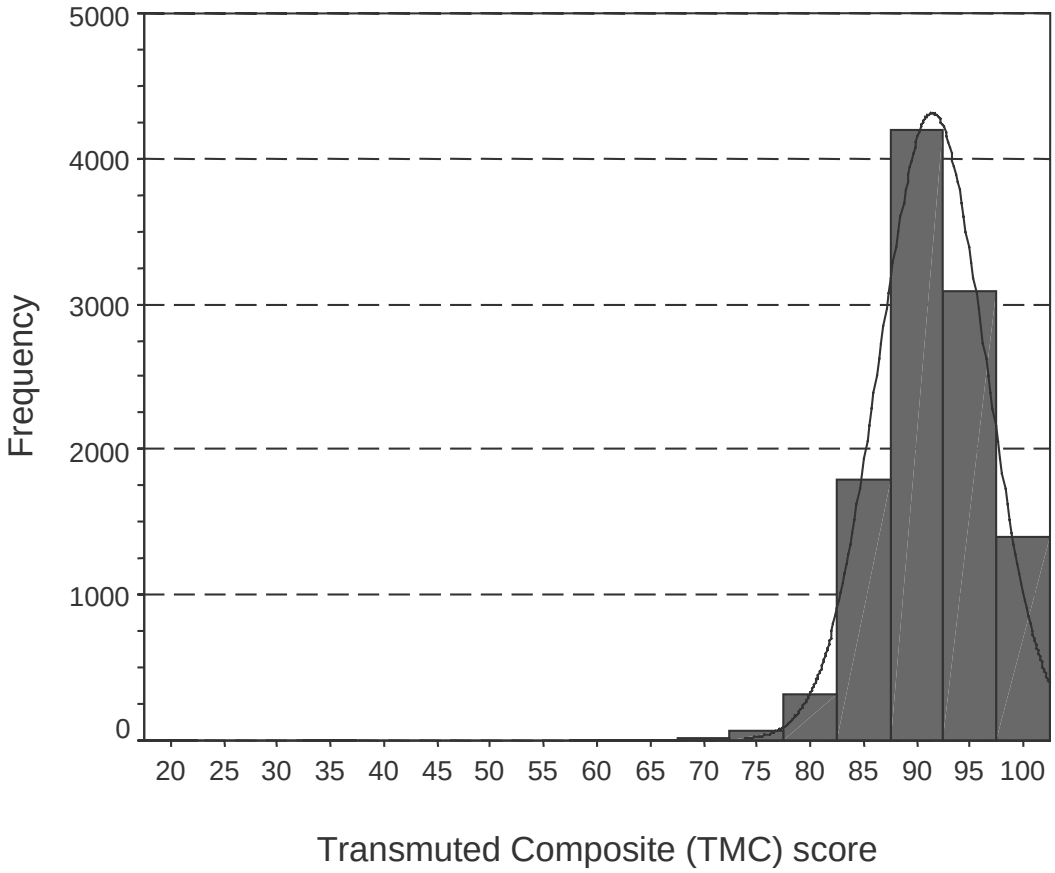
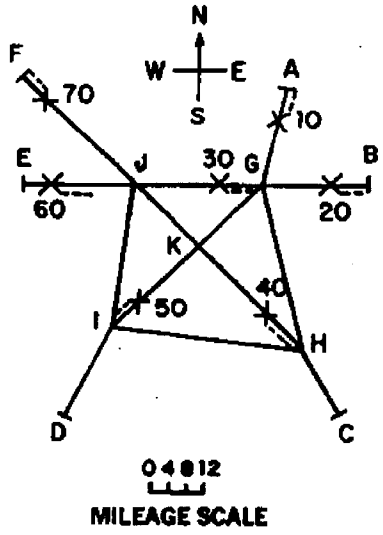


Figure 2: Frequency distribution of TMC scores in competitively hired ATCS sample: *N* = 10,869 **(** $\overline{X} = 91.46, \sigma = 5.02$ **)**

AIRCRAFT	ALTITUDE	SPEED	ROUTE
10	7000	480	AGKHC
20	7000	480	BGJE
30	7000	240	AGJE
40	6500	240	CHKJF
50	6500	240	DIKGB
60	8000	480	DIKJE
70	8000	480	FJKID



SAMPLE QUESTION

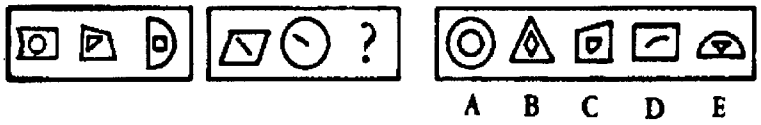
WHICH AIRCRAFT WILL CONFLICT?

A. 60 AND 70
B. 40 AND 70
C. 20 AND 30
D. NONE OF THESE

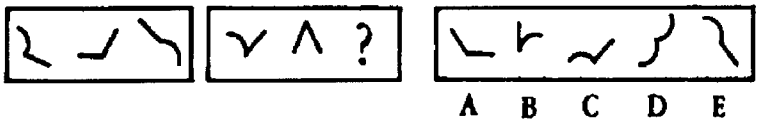
Figure 3: Example Multiplex Controller Aptitude Test (MCAT) item

Symbols

1.



2.



Letters

1) XCXDXEX A) FX B) FG C) XF D) EF E) XG

2) ARCSETG A) HI B) HU C) UJ D) UI E) IV

Figure 4: Example Abstract Reasoning (ABSR) item

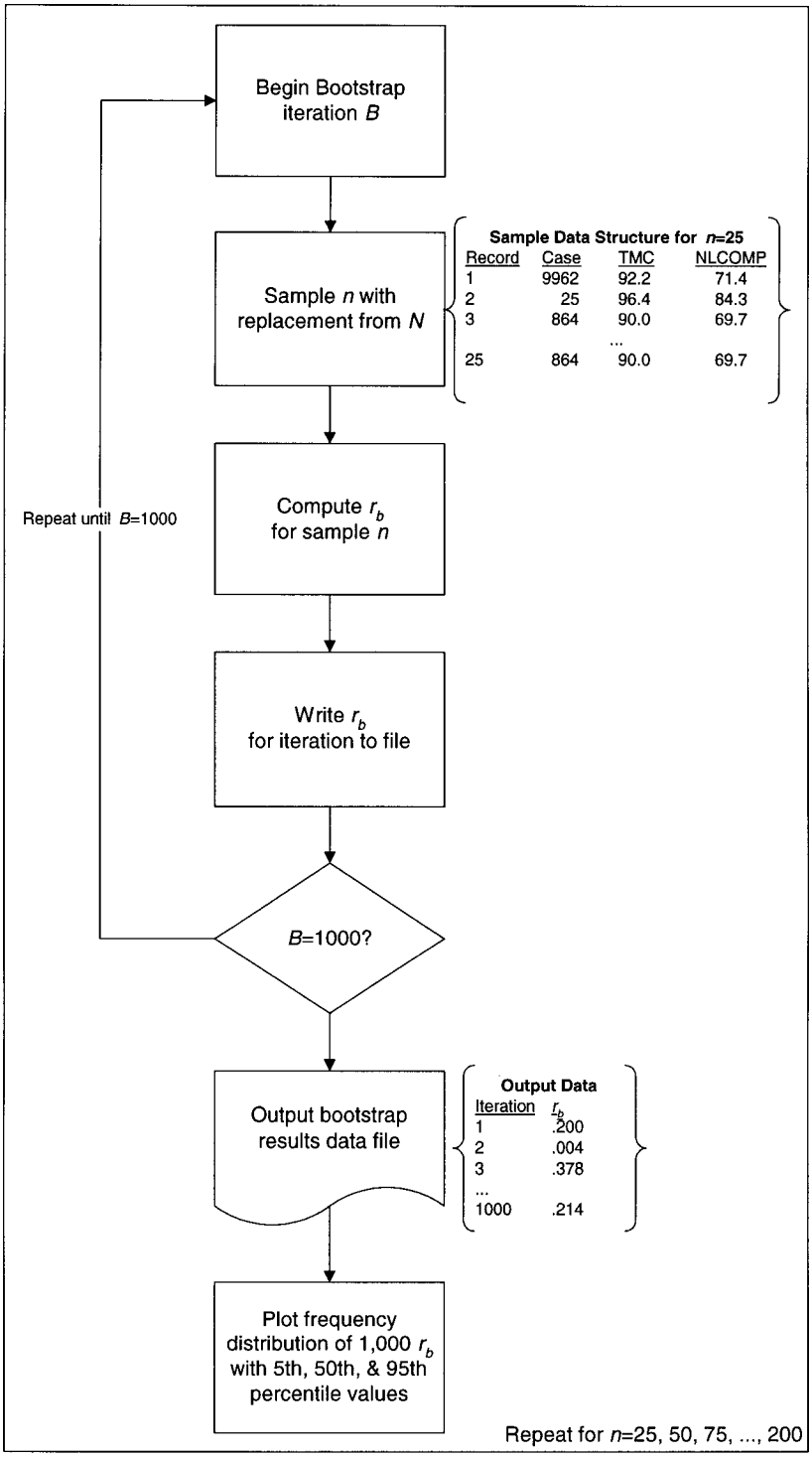


Figure 5: Flowchart of bootstrap analysis of simple, uncorrected TMC-NLCOMP correlations

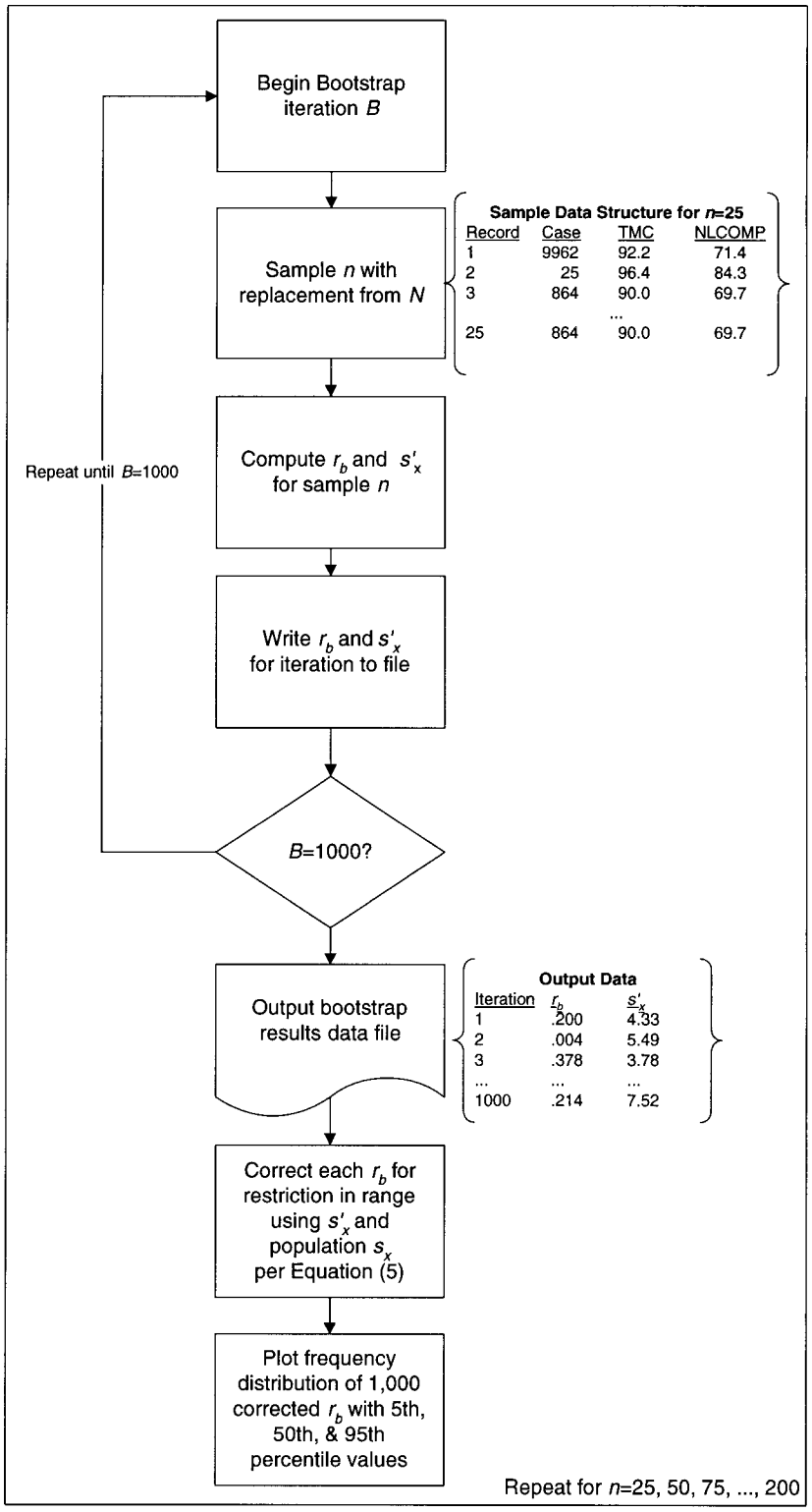


Figure 6: Flowchart of bootstrap analysis of TMC-NLCOMP correlations with corrections for restriction in range

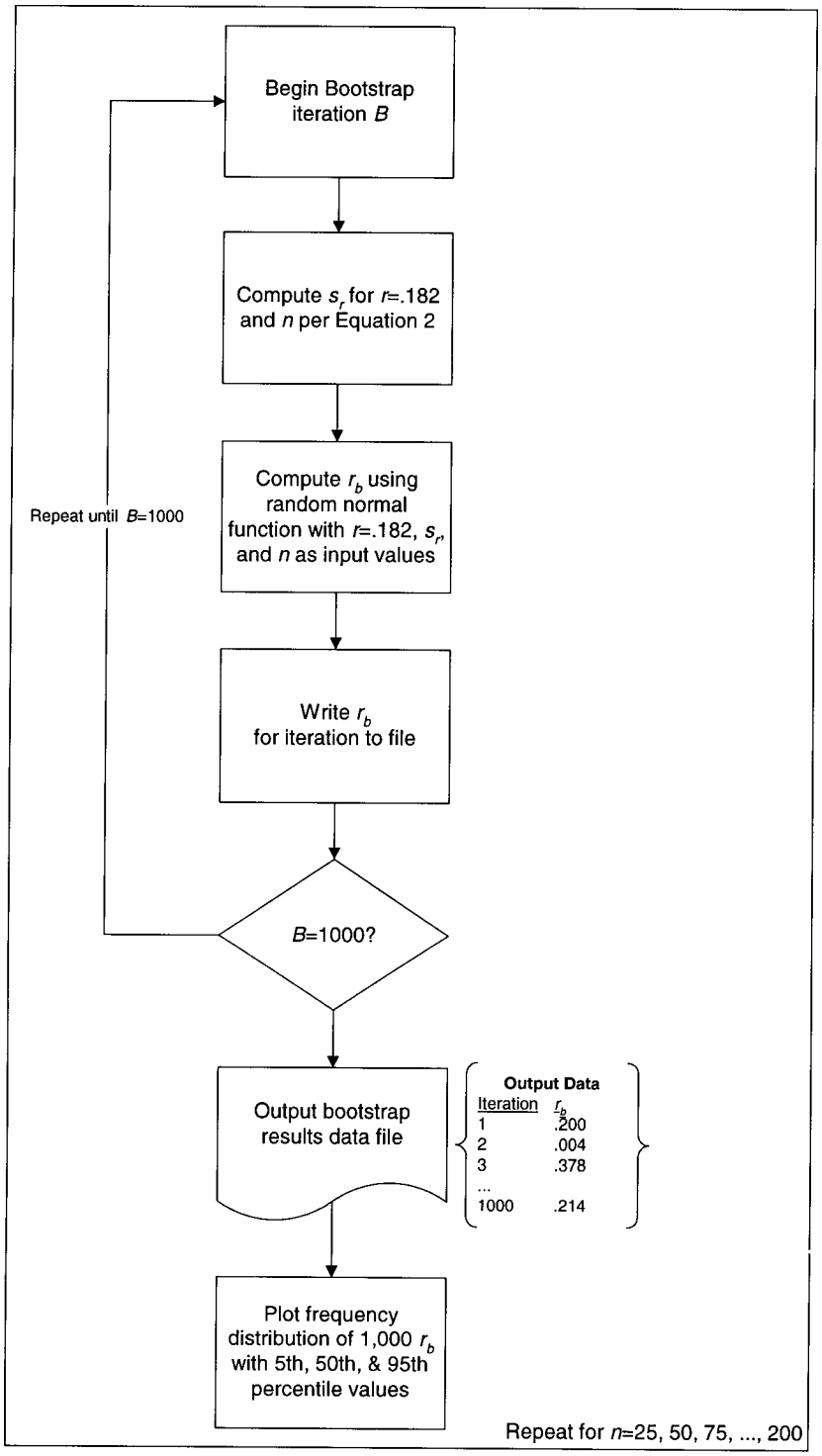


Figure 7: Flowchart of bootstrap analysis of correlations from bivariate normal population where $\rho = .182$ and σ computed by Equation 2 for bootstrap sample size of *n*

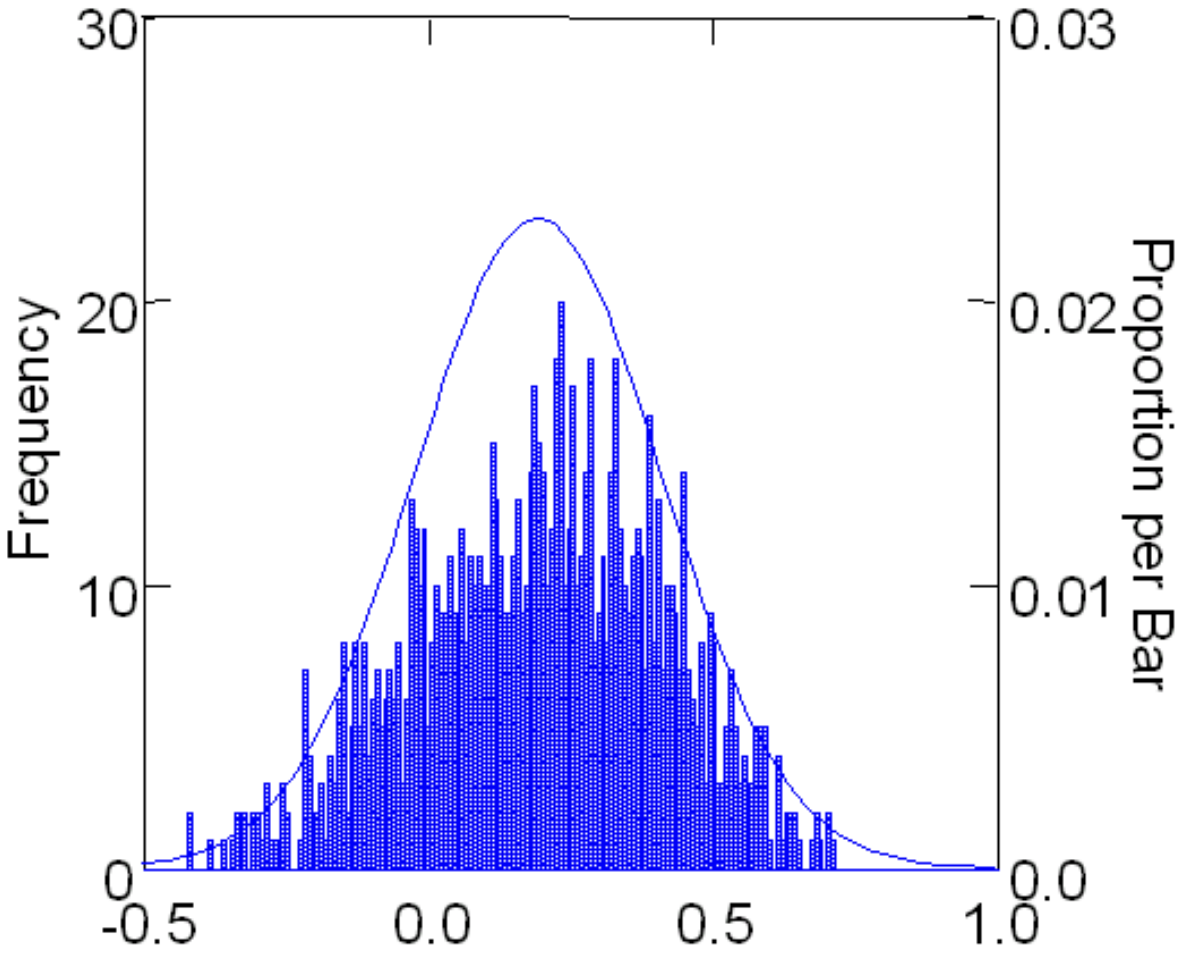
TABLE

Table 1
Estimates of 90% confidence interval and median for uncorrected and corrected TMC-NLCOMP validity coefficients (r_b) for B = 1,000 bootstrap samples of n = 25, 50, ..., 200

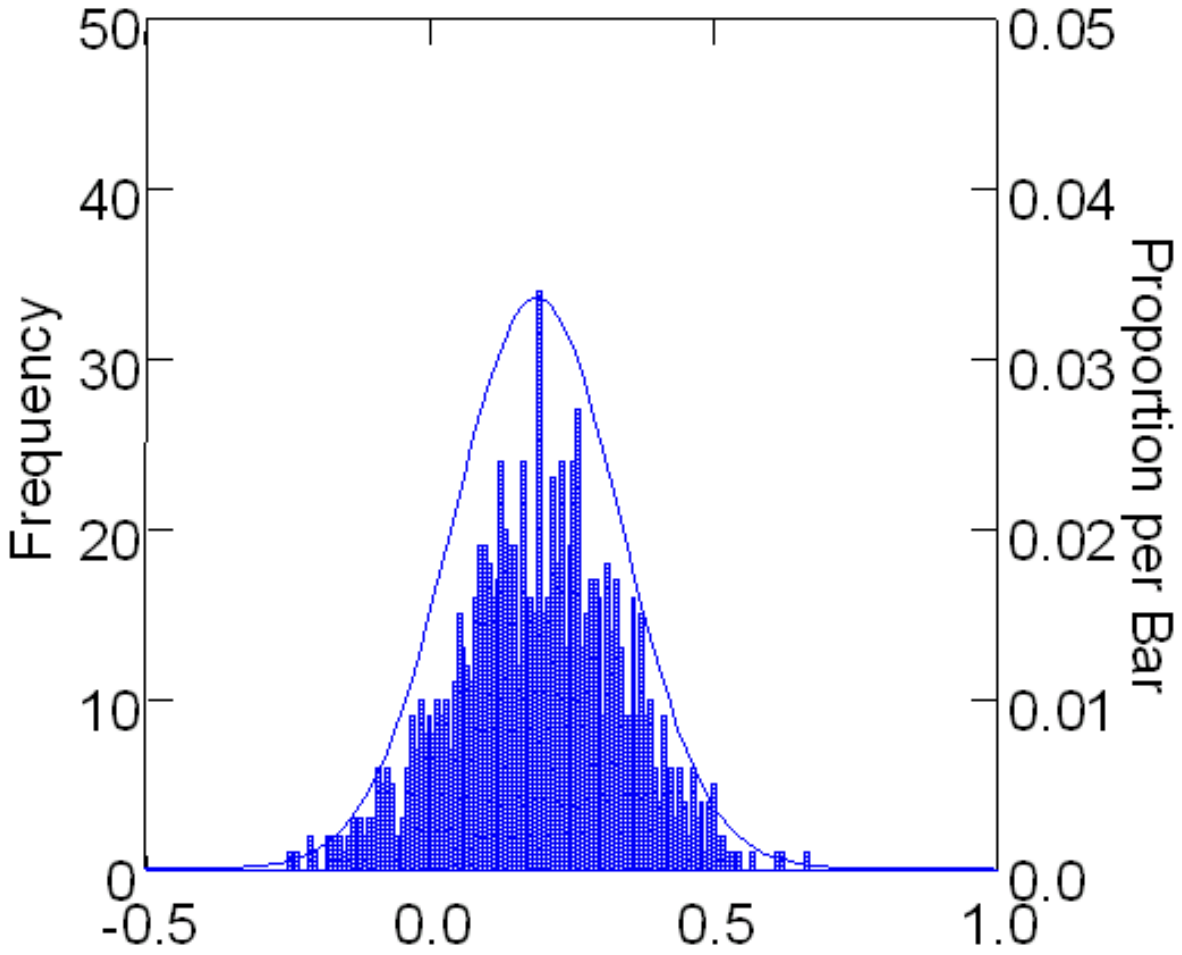
	90% CI Boundaries and Median		
Analysis	5%	50%	95%
	n = 25		
Uncorrected	-0.251	0.207	0.587
Corrected	-0.589	0.511	0.898
	n = 50		
Uncorrected	-0.110	0.195	0.476
Corrected	-0.298	0.487	0.836
	n = 75		
Uncorrected	-0.072	0.186	0.423
Corrected	-0.198	0.469	0.796
	n = 100		
Uncorrected	-0.020	0.195	0.404
Corrected	-0.055	0.488	0.779
	n = 125		
Uncorrected	-0.008	0.187	0.377
Corrected	-0.021	0.473	0.753
	n = 150		
Uncorrected	-0.001	0.191	0.368
Corrected	-0.003	0.481	0.744
	n = 175		
Uncorrected	0.016	0.193	0.346
Corrected	0.046	0.483	0.719
	n = 200		
Uncorrected	0.034	0.190	0.332
Corrected	0.094	0.477	0.703

APPENDIX A

Distributions of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 25, 50, …, 200

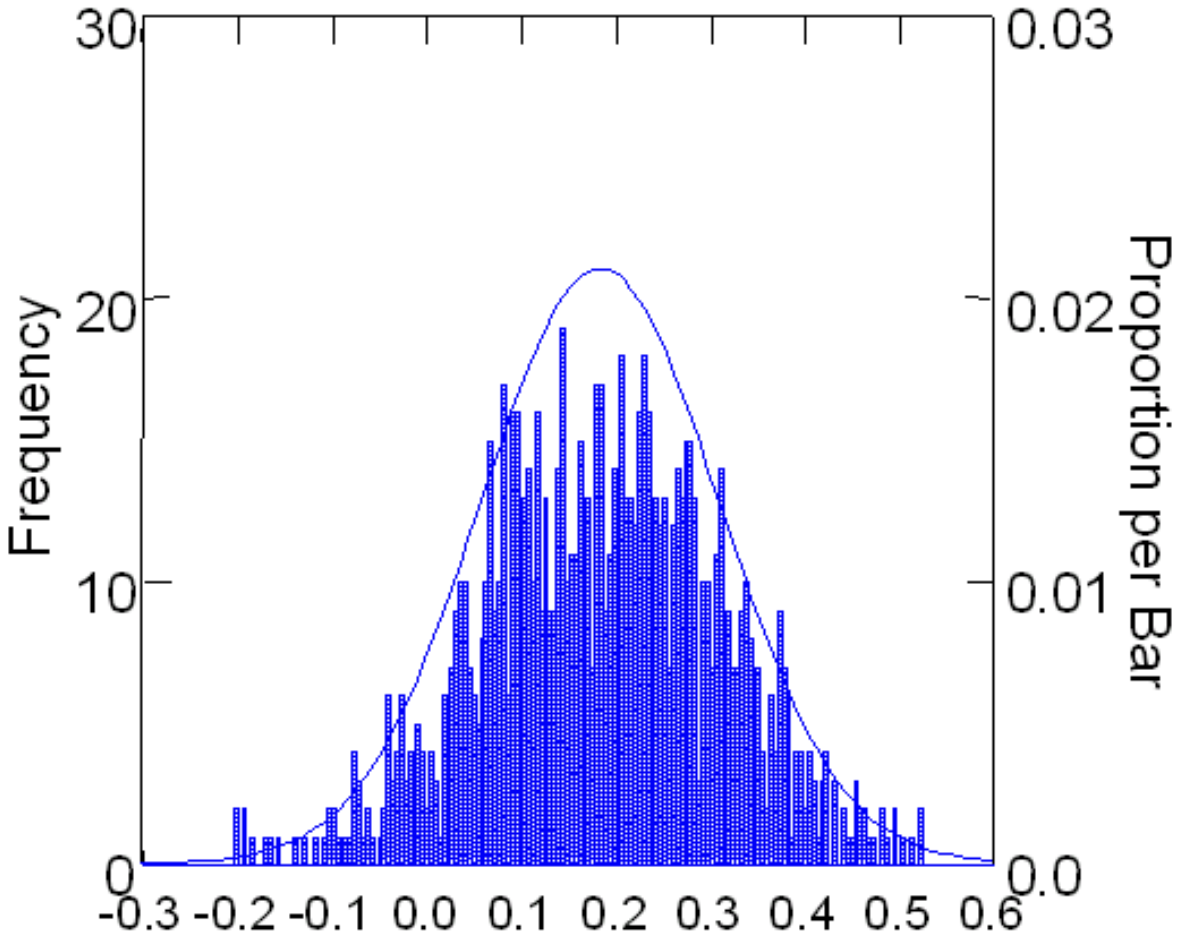


Graph A-1: Distribution of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 25



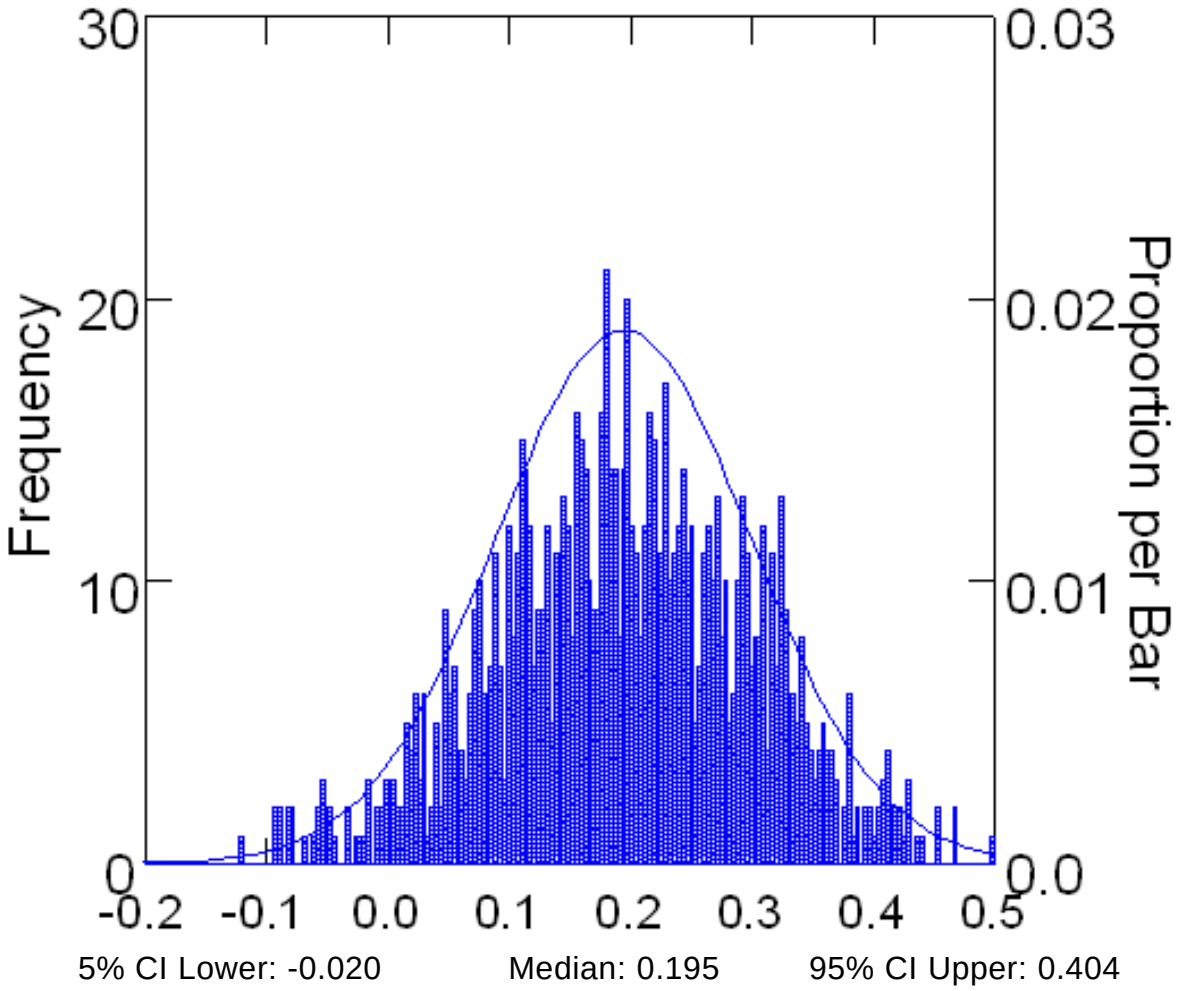
5% CI Lower: -0.110 Median: 0.195 95% CI Upper: 0.476

Graph A-2: Distribution of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 50



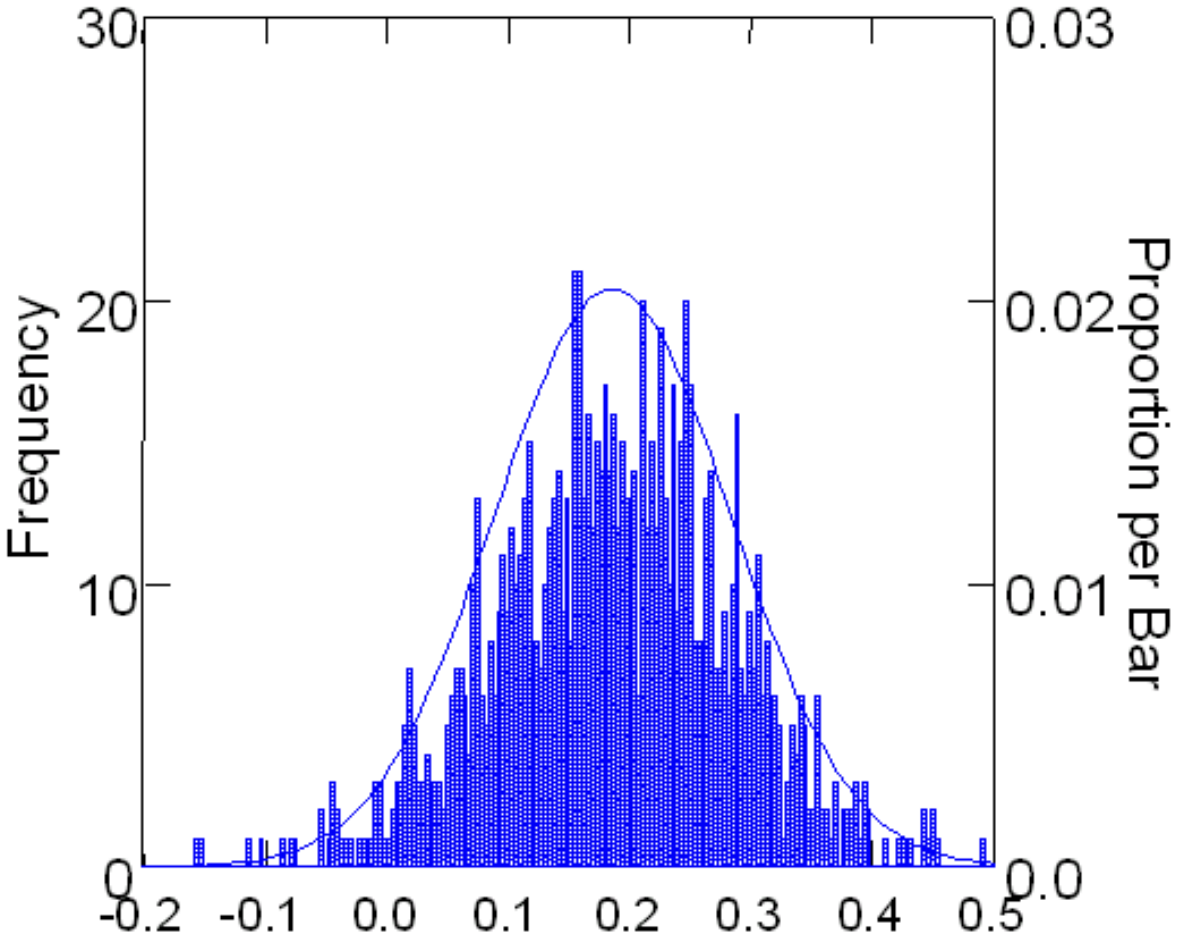
5% CI Lower: -0.072 Median: 0.186 95% CI Upper: 0.423

Graph A-3: Distribution of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 75



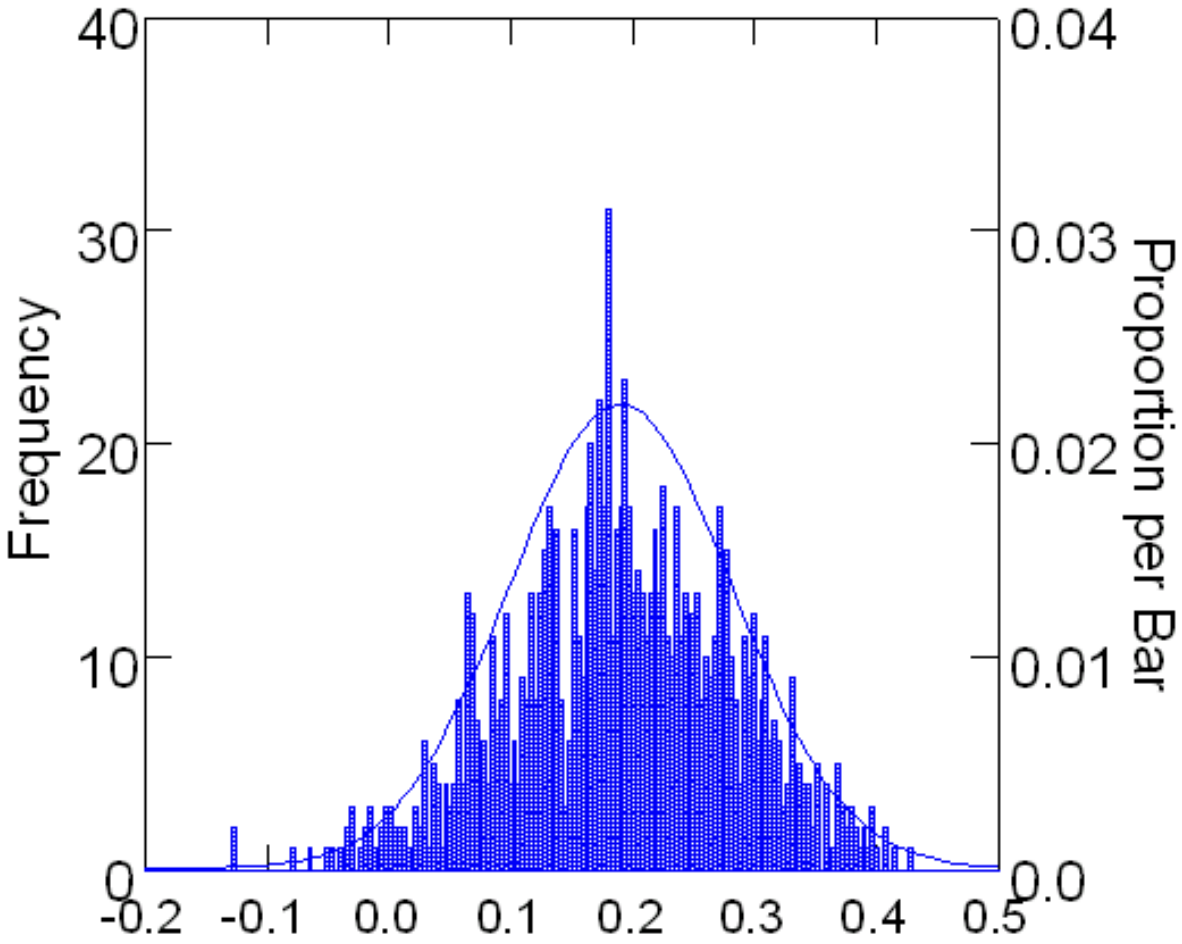
5% CI Lower: -0.020 Median: 0.195 95% CI Upper: 0.404

Graph A-4: Distribution of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 100



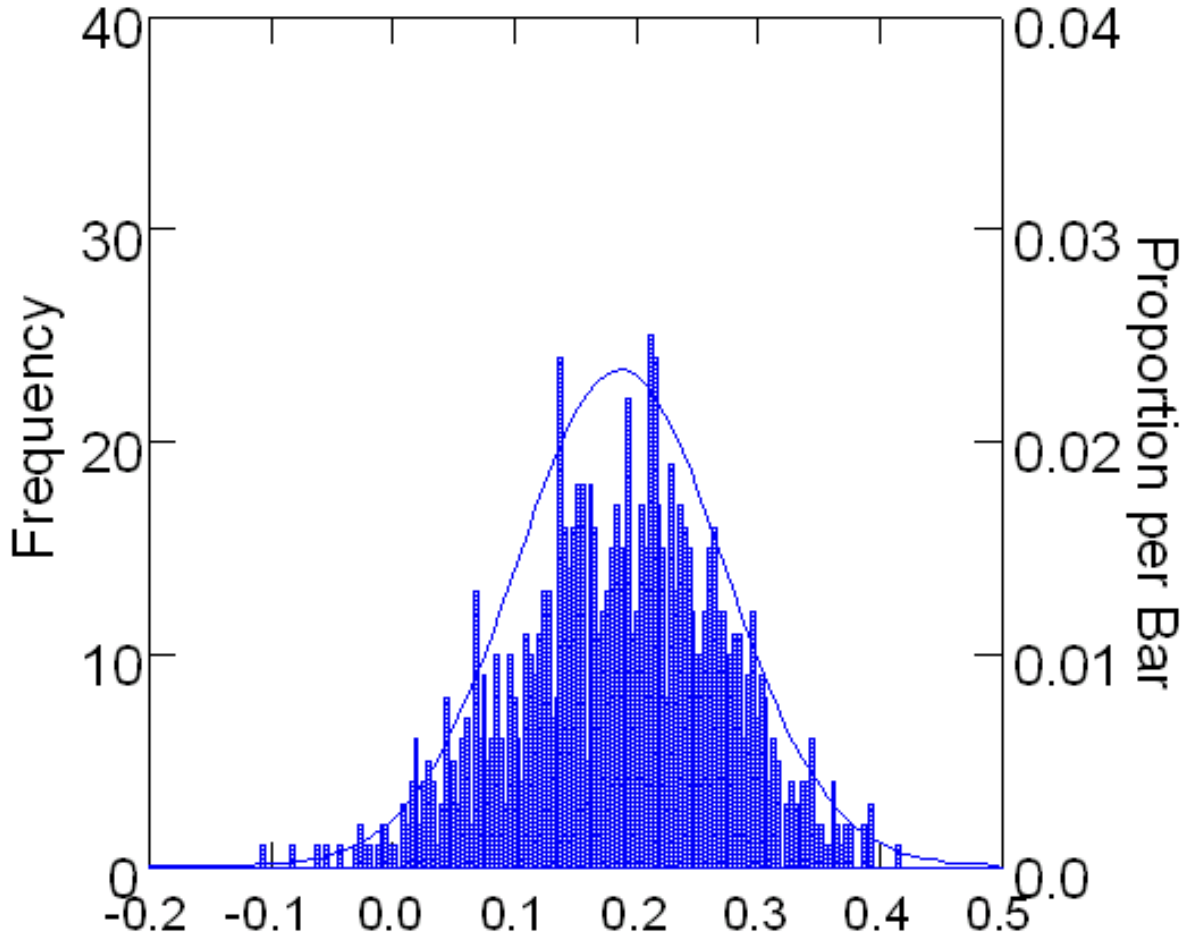
5% CI Lower: -0.008 Median: 0.187 95% CI Upper: 0.377

Graph A-5: Distribution of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 125



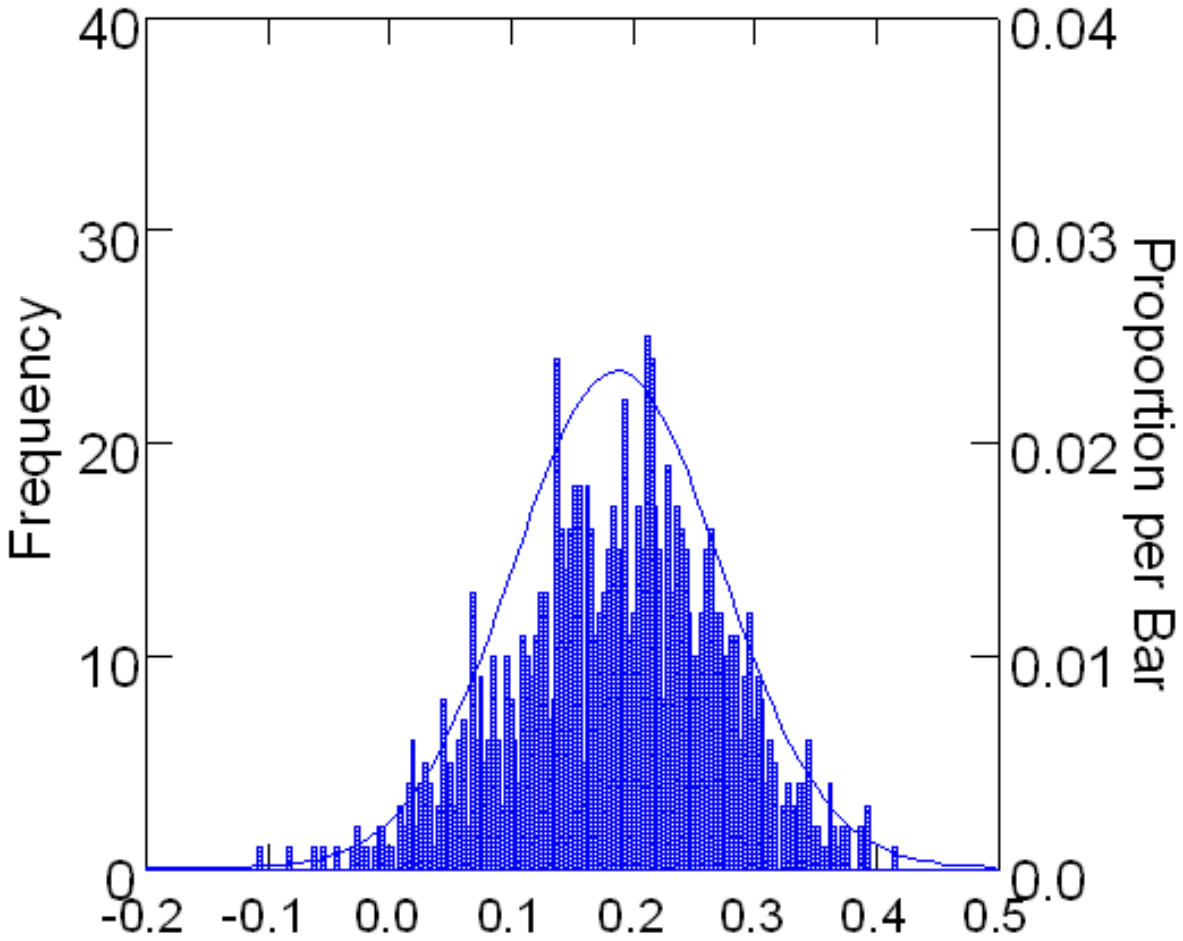
5% CI Lower: -0.001 Median: 0.191 95% CI Upper: 0.368

Graph A-6: Distribution of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 150



5% CI Lower: 0.016 Median: 0.193 95% CI Upper: 0.346

Graph A-7: Distribution of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 175

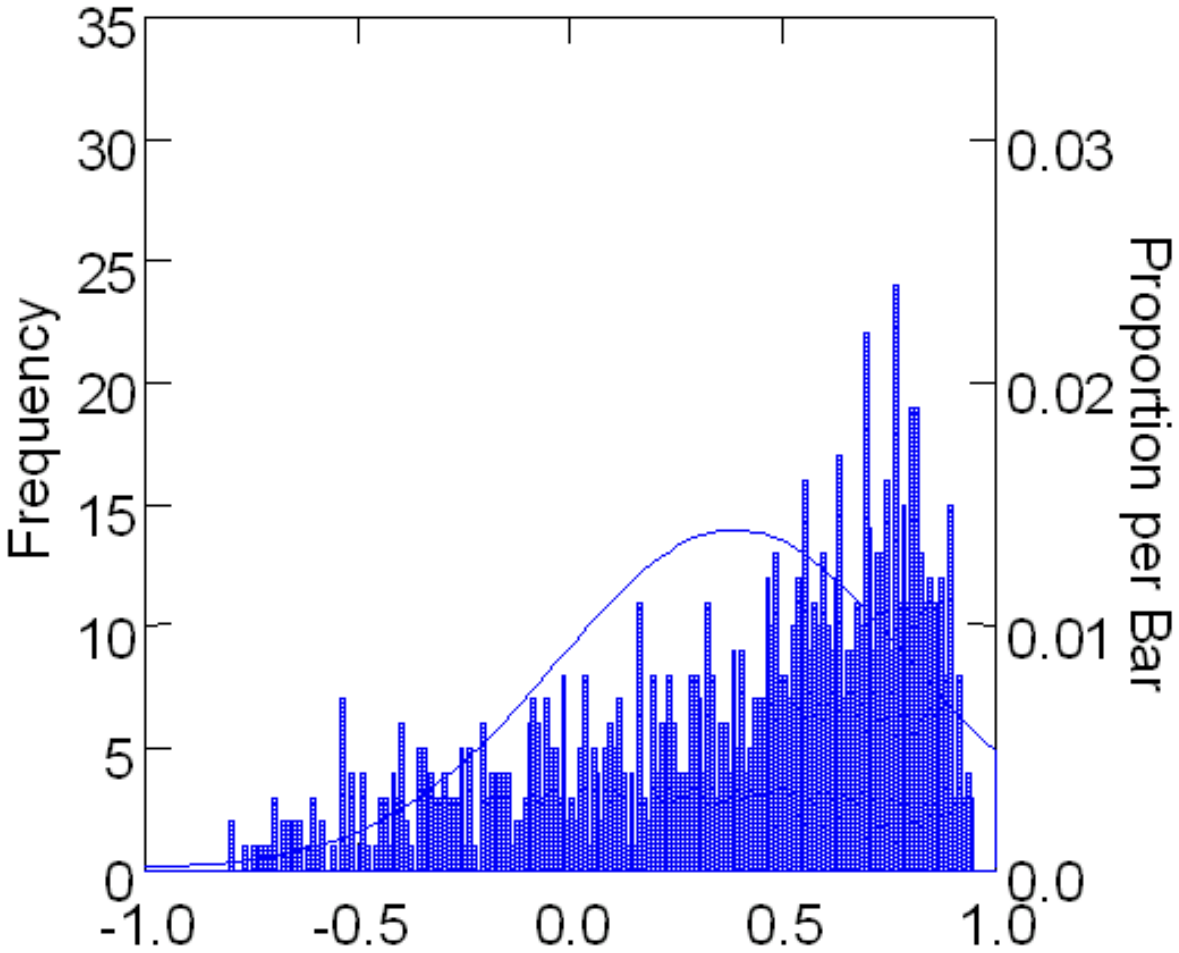


5% CI Lower: 0.034 Median: 0.190 95% CI Upper: 0.332

Graph A-8: Distribution of TMC-NLCOMP correlations uncorrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 200

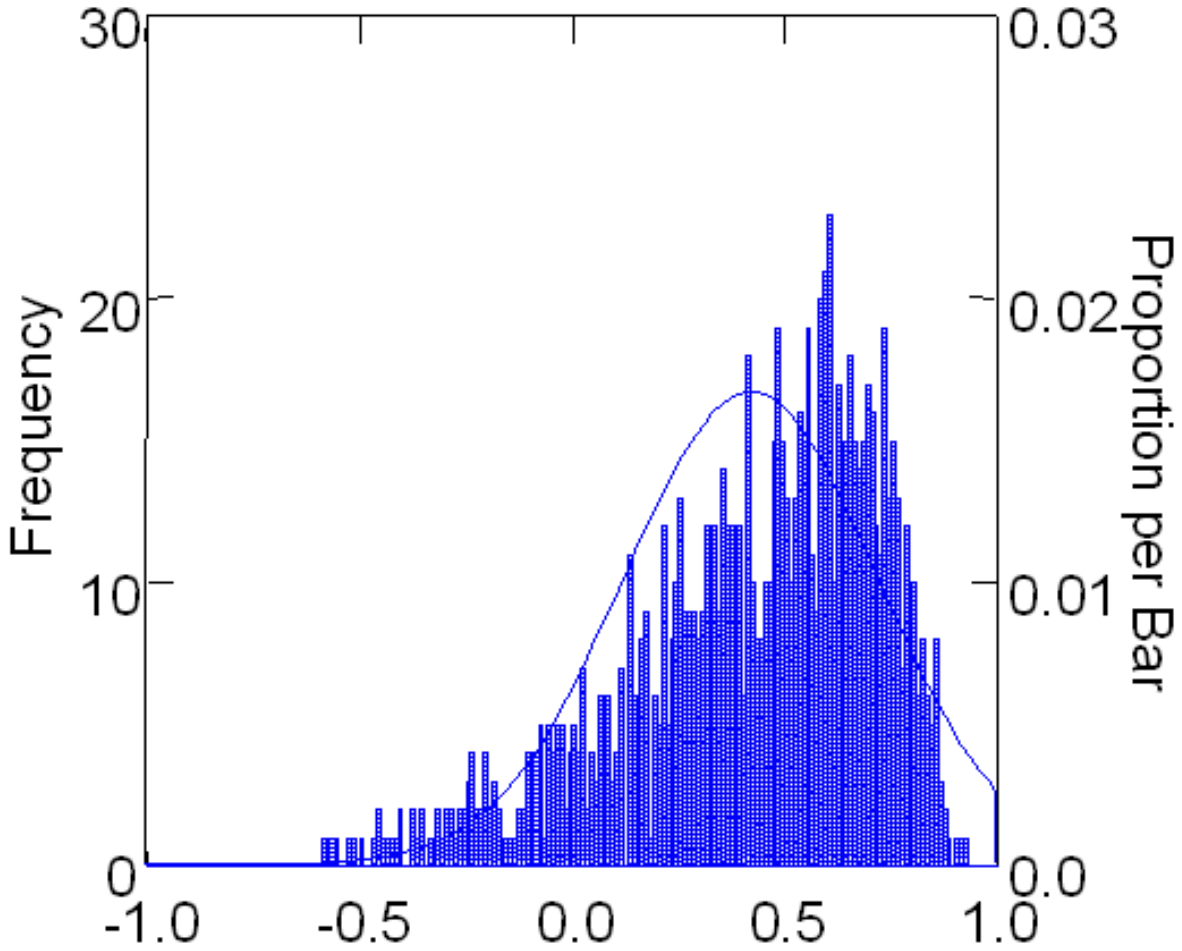
APPENDIX B

Distributions of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 25, 50, …, 200



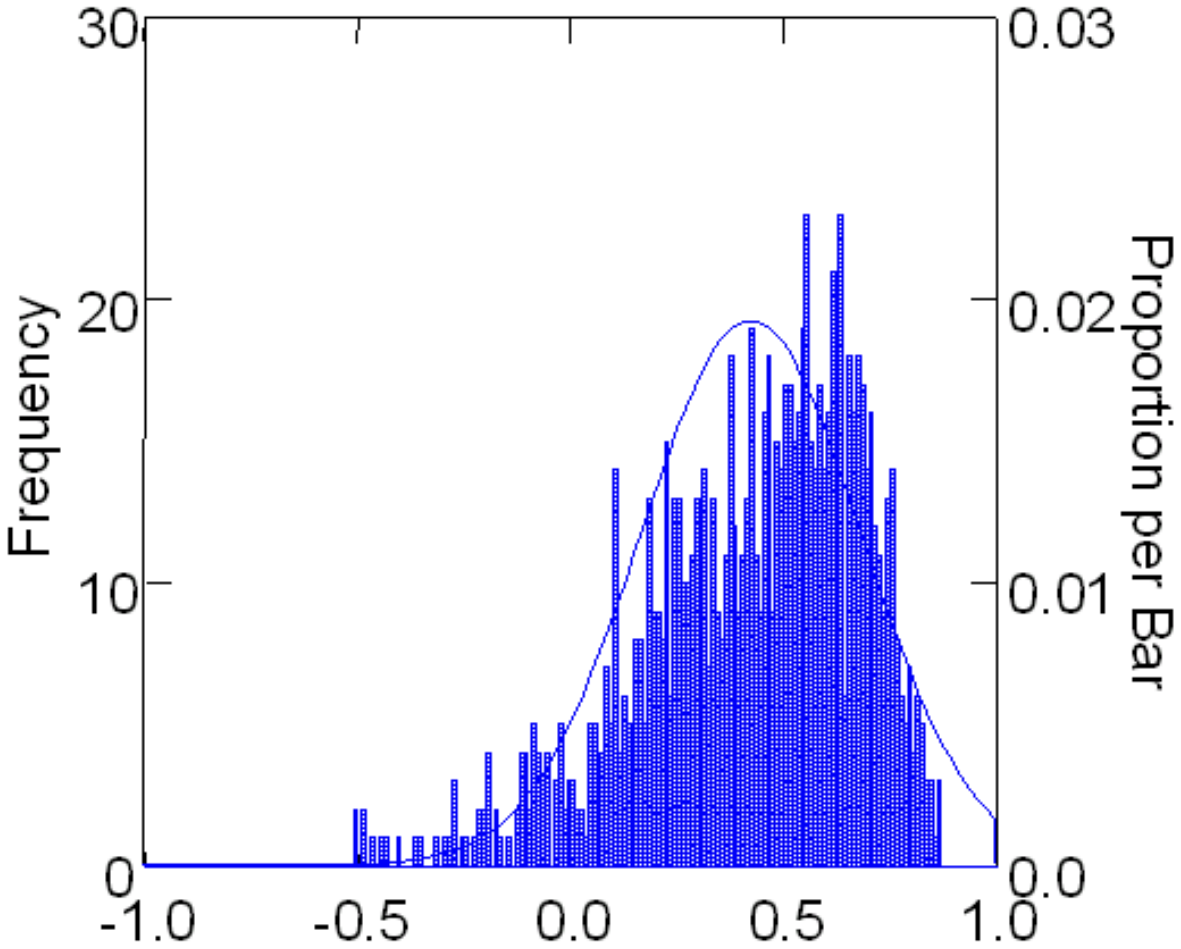
5% CI Lower: -0.589 Median: 0.511 95% CI Upper: 0.898

Graph B-1: Distribution of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 25



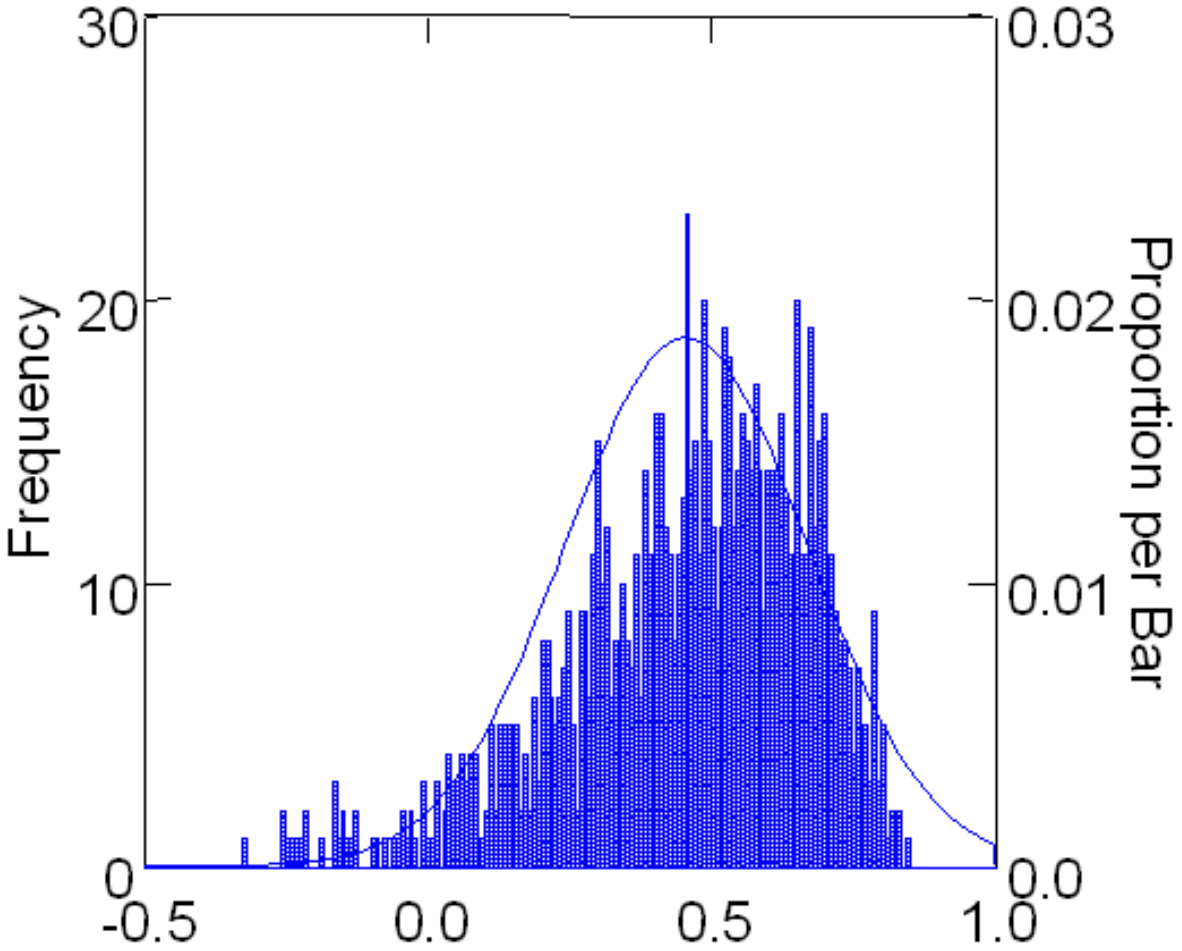
5% CI Lower: -0.298 Median: 0.487 95% CI Upper: 0.836

Graph B-2: Distribution of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 50

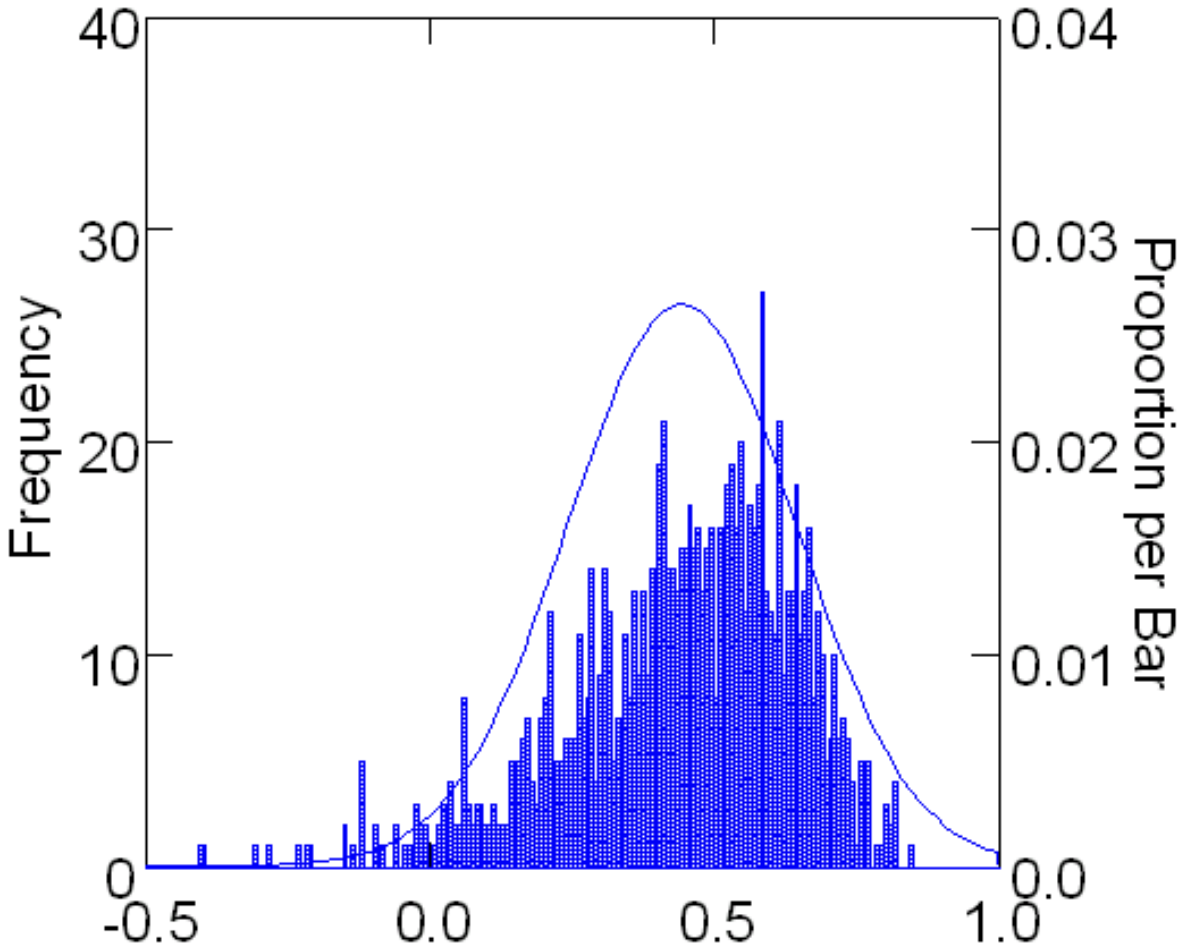


5% CI Lower: -0.198 Median: 0.469 95% CI Upper: 0.796

Graph B-3: Distribution of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 75

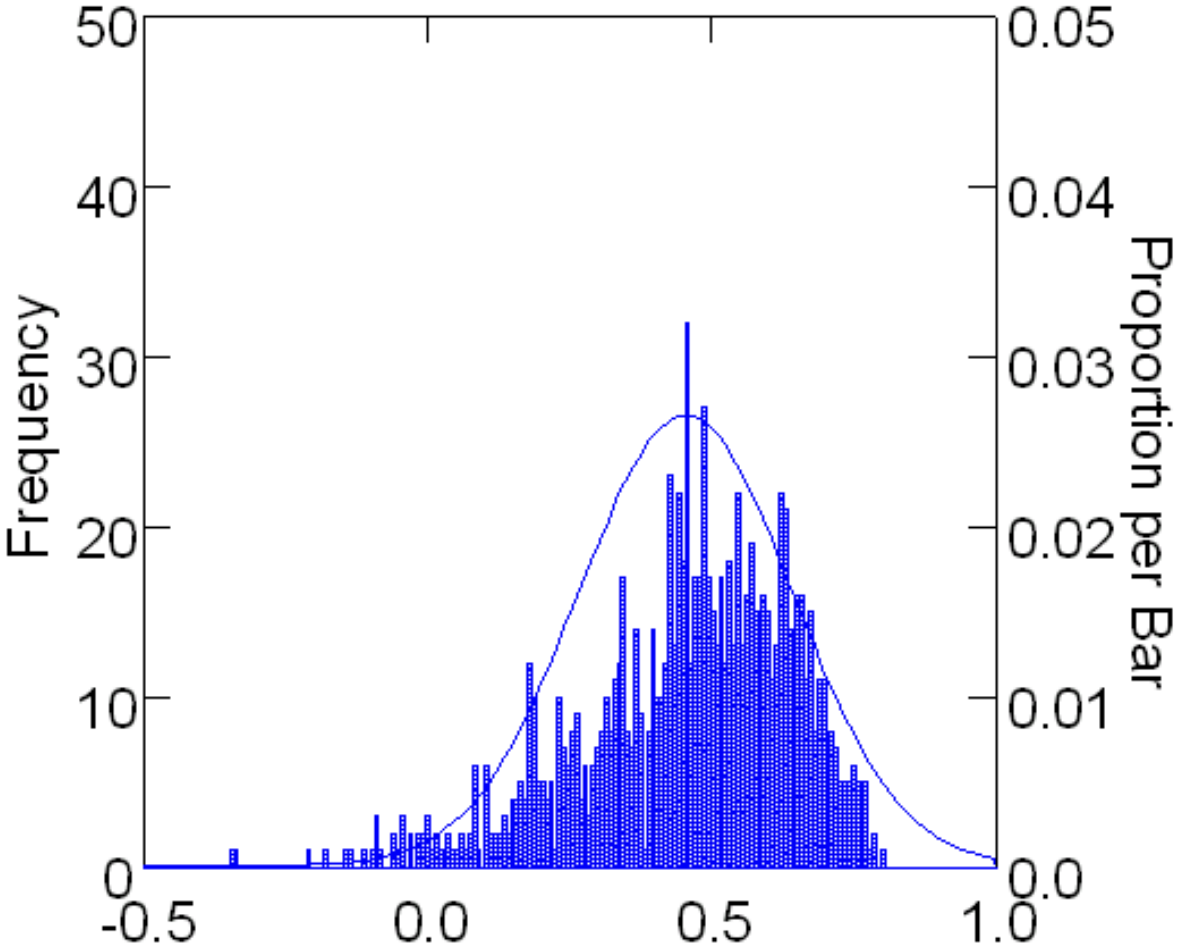


Graph B-4: Distribution of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 100



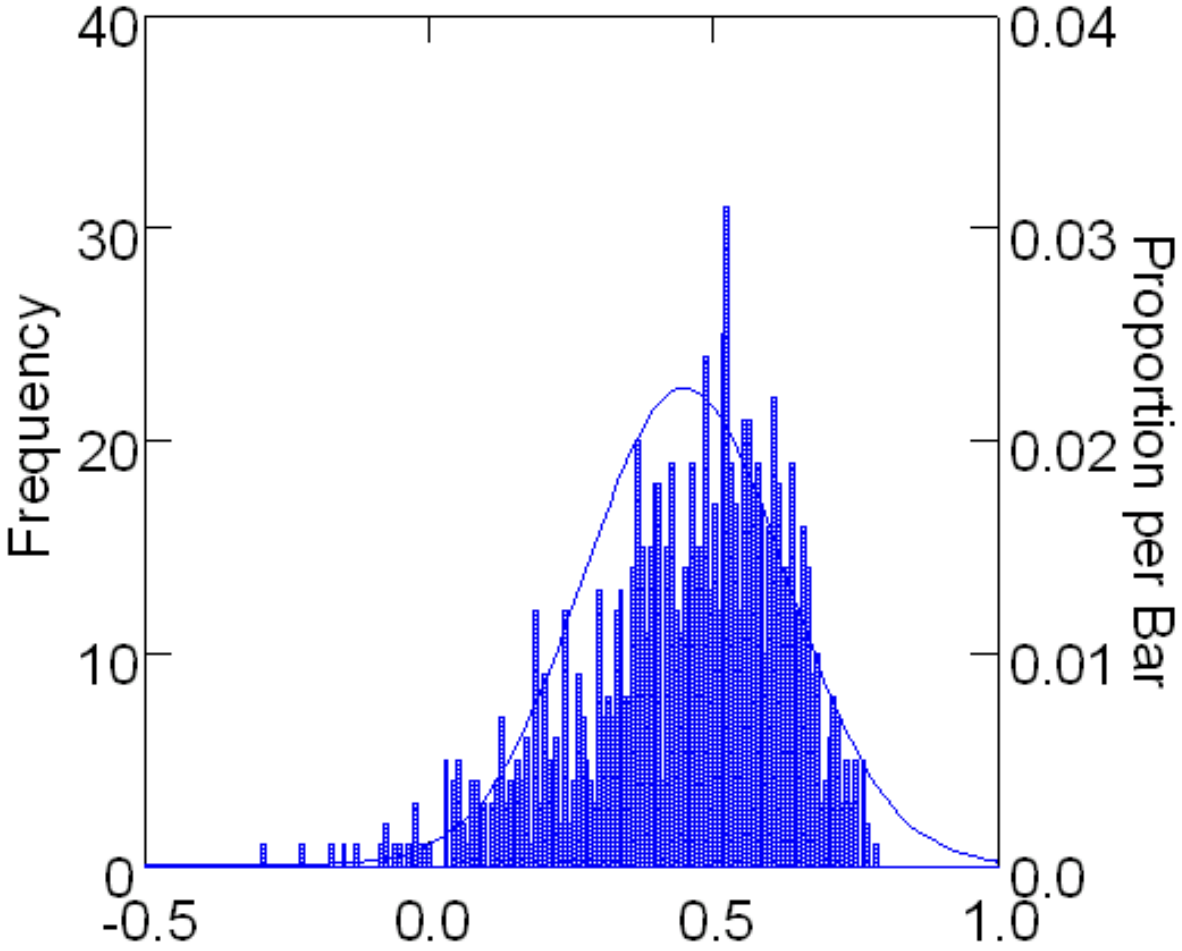
5% CI Lower: -0.021 Median: 0.473 95% CI Upper: 0.753

Graph B-5: Distribution of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 125



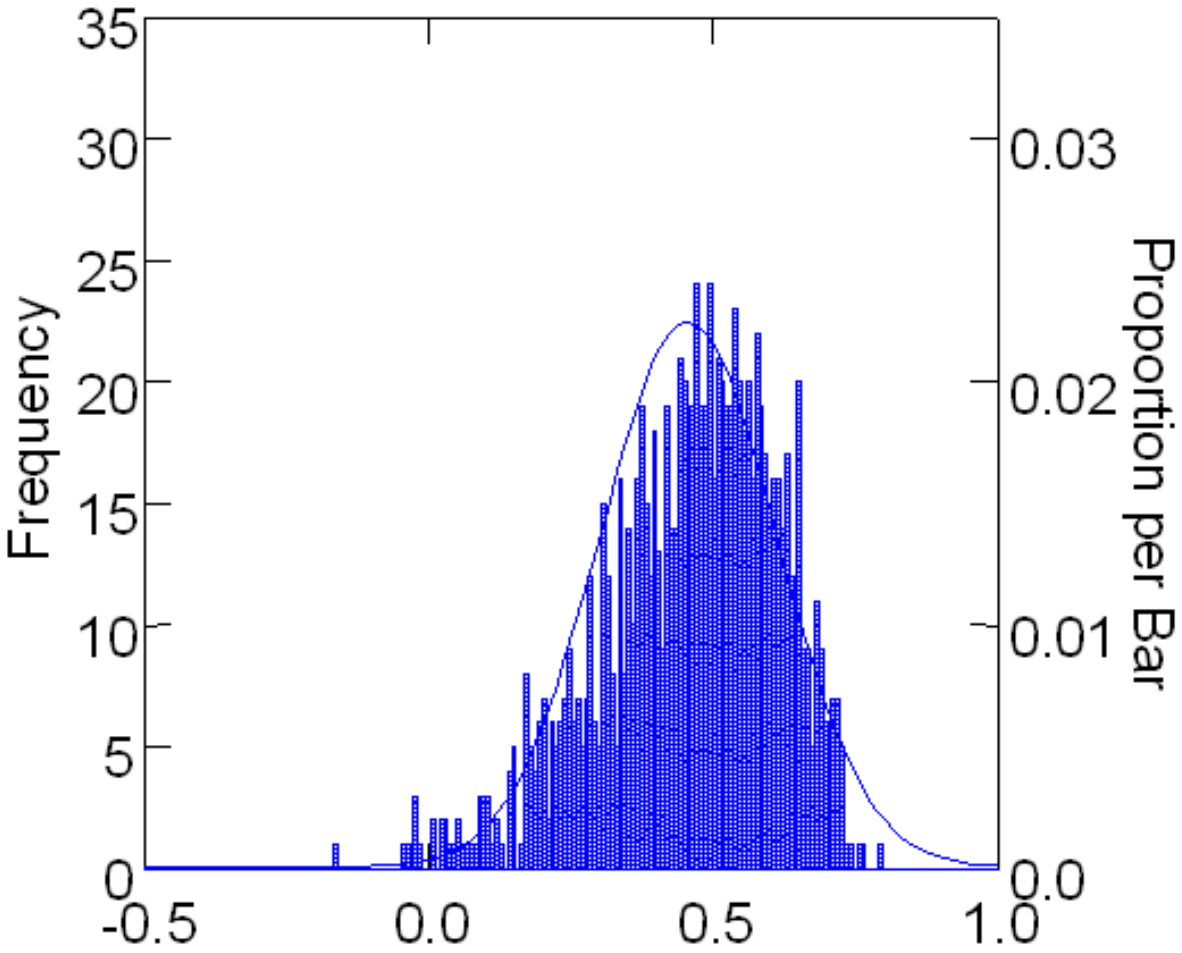
5% CI Lower: -0.003 Median: 0.481 95% CI Upper: 0.744

Graph B-6: Distribution of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 150



5% CI Lower: 0.046 Median: 0.483 95% CI Upper: 0.719

Graph B-7: Distribution of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 175

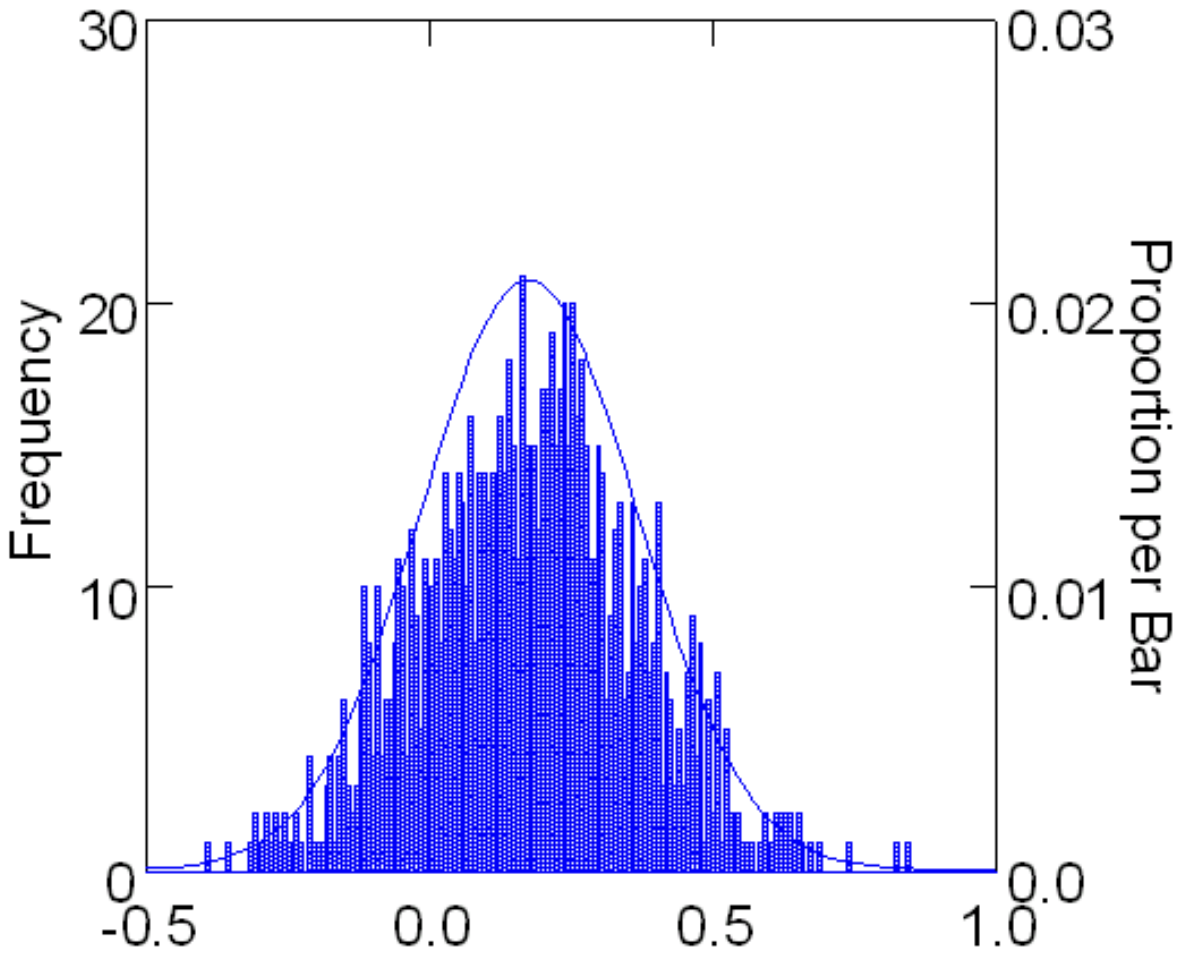


5% CI Lower: 0.094 Median: 0.477 95% CI Upper: 0.703

Graph B-8: Distribution of TMC-NLCOMP correlations corrected for range restriction for *B* = 1,000 bootstrap samples of *n* = 200

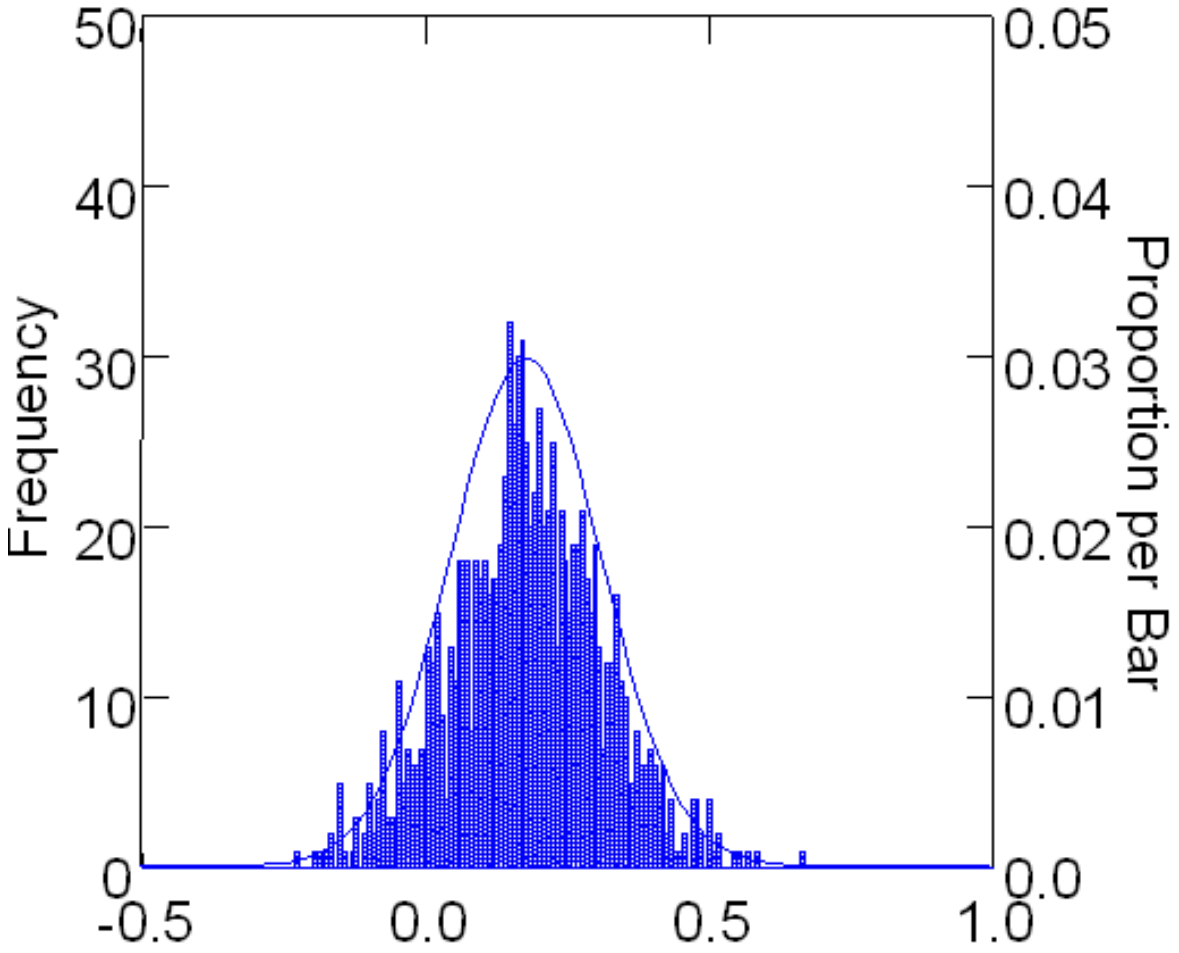
APPENDIX C

Distributions of TMC-NLCOMP correlations generated for a bivariate normal population with parameters ρ and S_R from B = 1,000 bootstrap samples of n = 25, 50, …, 200



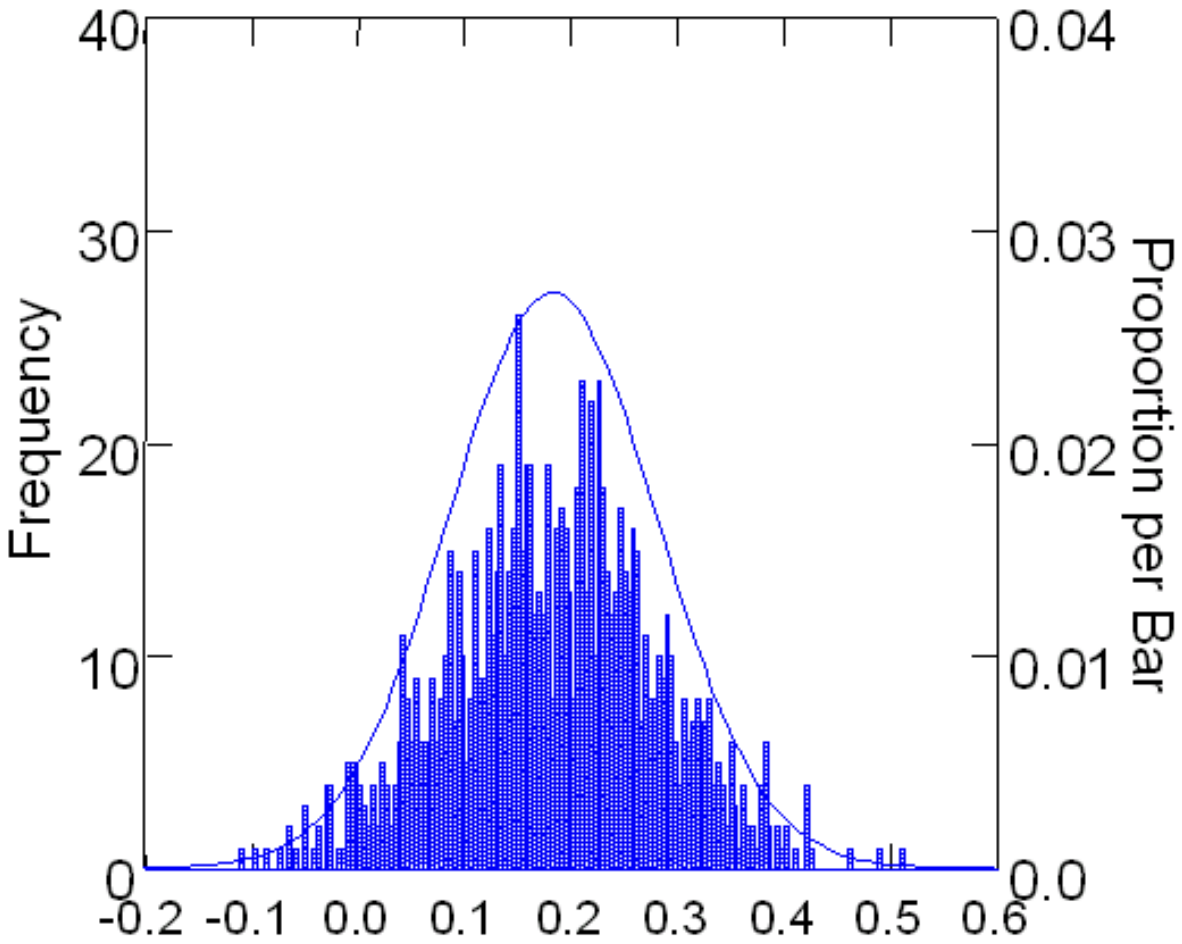
5% CI Lower: -0.193 ρ = .182, s_r = .1974 95% CI Upper: 0.545

Graph C-1: Distribution of TMC-NLCOMP correlations generated for a bivariate normal population with parameters ρ and s_r from B = 1,000 bootstrap samples of n = 25



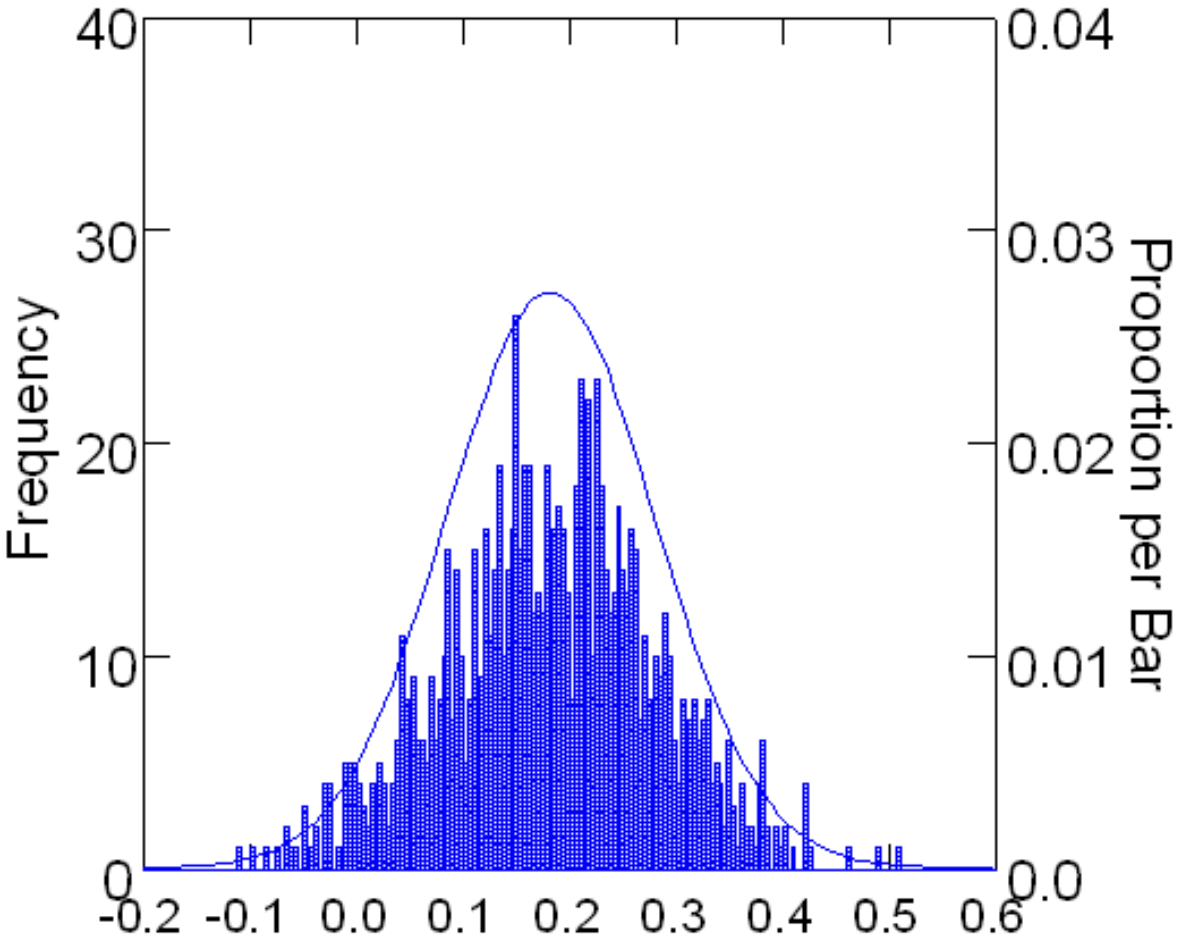
5% CI Lower: -0.088 ρ = .182, s_r = .1381 95% CI Upper: 0.442

Graph C-2: Distribution of TMC-NLCOMP correlations generated for a bivariate normal population with parameters ρ and s_r from B = 1,000 bootstrap samples of n = 50



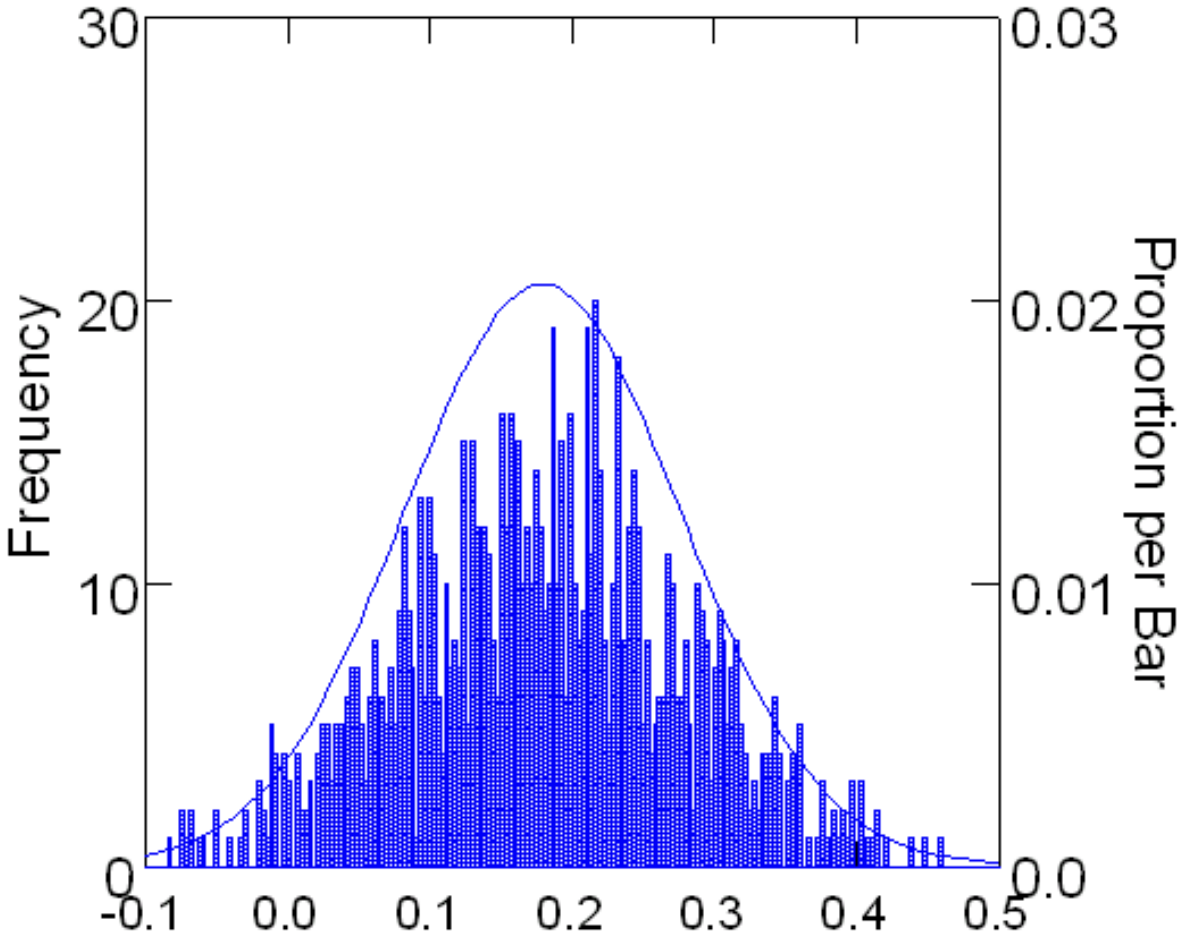
5% CI Lower: -0.012 ρ = .182, s_r = .1124 95% CI Upper: 0.377

Graph C-3: Distribution of TMC-NLCOMP correlations generated for a bivariate normal population with parameters ρ and s_r from B = 1,000 bootstrap samples of n = 75



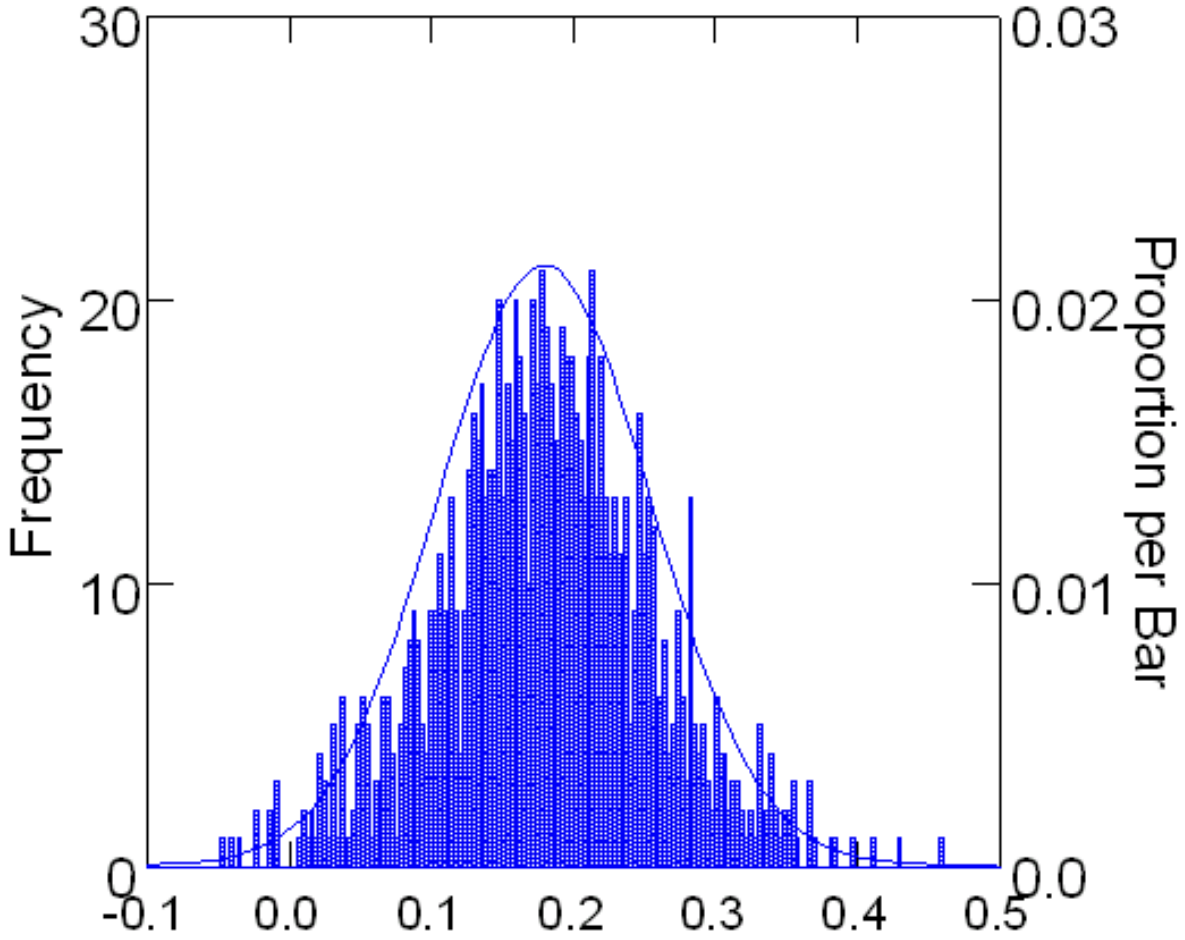
5% CI Lower: -0.010 ρ = .182, s_r = .0972 95% CI Upper: 0.360

Graph C-4: Distribution of TMC-NLCOMP correlations generated for a bivariate normal population with parameters and s_r from B = 1,000 bootstrap samples of n = 100



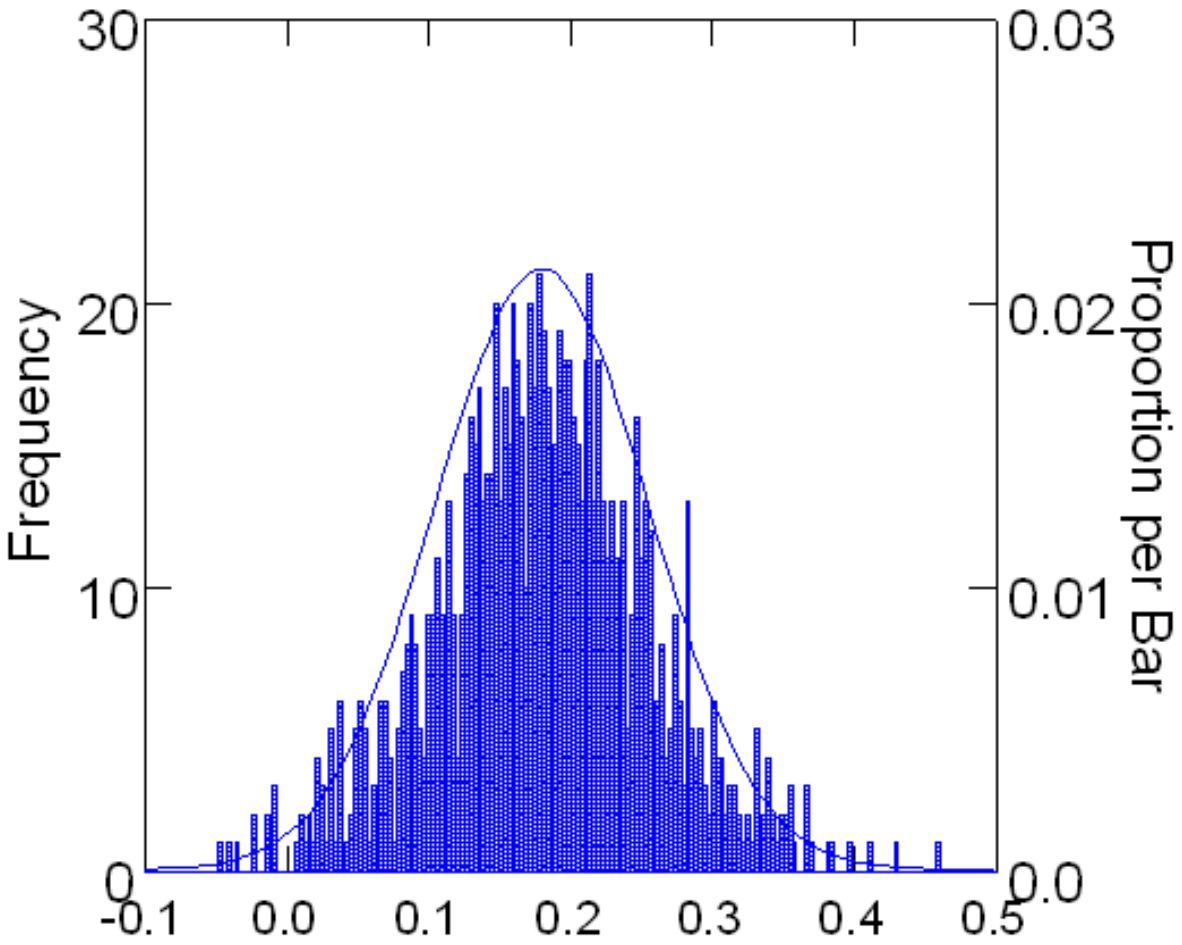
5% CI Lower: 0.011 ρ = .182, s_r = .0868 95% CI Upper: 0.373

Graph C-5: Distribution of TMC-NLCOMP correlations generated for a bivariate normal population with parameters ρ and s_r from B = 1,000 bootstrap samples of n = 125



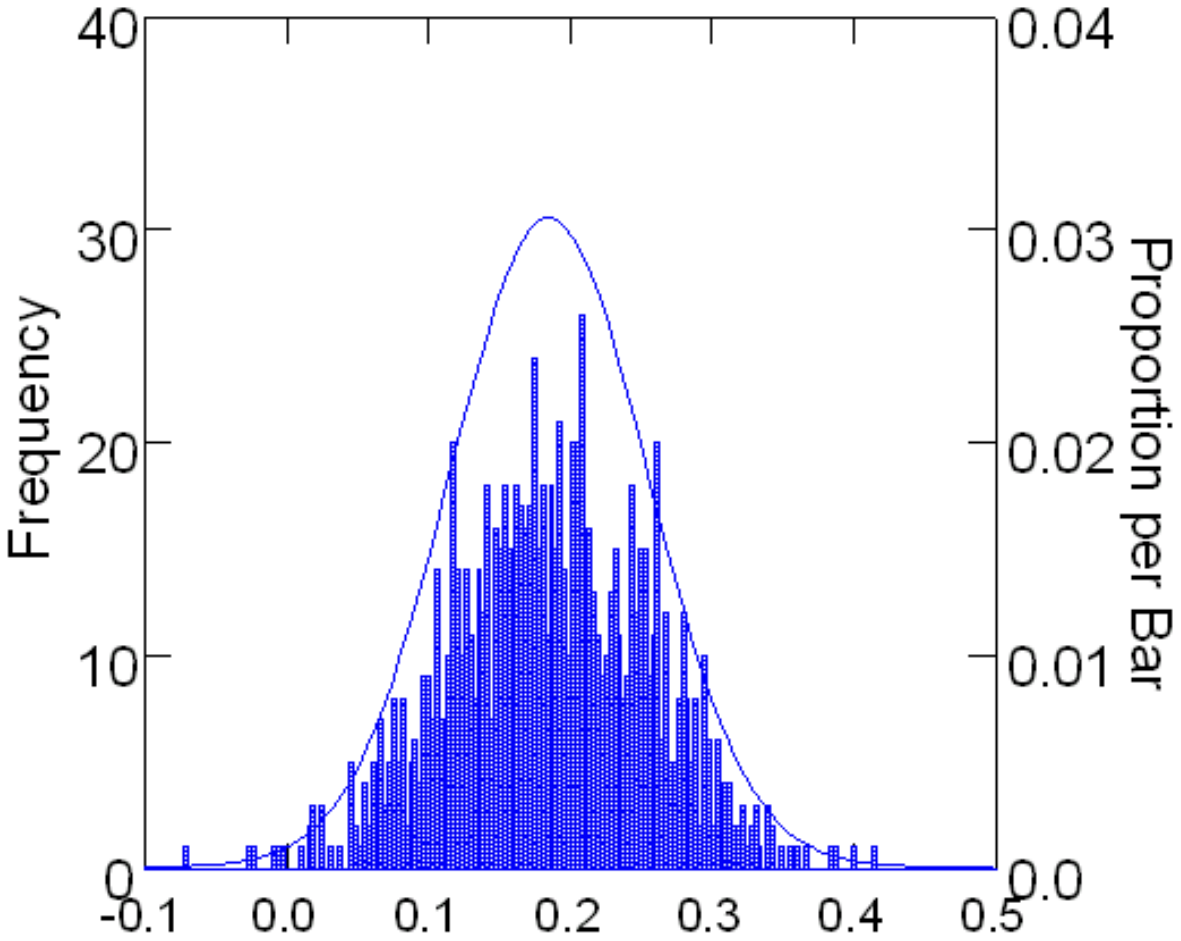
5% CI Lower: 0.025 ρ = .182, s_r = .0792 95% CI Upper: 0.335

Graph C-6: Distribution of TMC-NLCOMP correlations generated for a bivariate normal population with parameters ρ and s_r from B = 1,000 bootstrap samples of n = 150



5% CI Lower: 0.026 ρ = .182, s_r = .0733 95% CI Upper: 0.337

Graph C-7: Distribution of TMC-NLCOMP correlations generated for a bivariate normal population with parameters ρ and s_r from B = 1,000 bootstrap samples of n = 175



5% CI Lower: 0.051 ρ = .182, s_r = .0685 95% CI Upper: 0.316

Graph C-8: Distribution of TMC-NLCOMP correlations generated for a bivariate normal population with parameters ρ and s_r from B = 1,000 bootstrap samples of n = 200