Journal of Modern Applied Statistical Methods

Volume 10 | Issue 2 Article 8

11-1-2011

Probabilistic Inferences for the Sample Pearson Product Moment Correlation

Jeffrey R. Harring
University of Maryland, harring@umd.edu

John A. Wasko
University of Maryland, john.wasko@us.army.mil

DOI: 10.22237/jmasm/1320120420

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 476-493

1538 – 9472/11/$95.00

Probabilistic Inferences for the Sample Pearson Product Moment Correlation

Jeffrey R. Harring John A. Wasko
University of Maryland,
College Park, MD

Fisher's correlation transformation is commonly used to draw inferences regarding the reliability of tests comprised of dichotomous or polytomous items. It is illustrated theoretically and empirically that omitting test length and difficulty results in inflated Type I error. An empirically unbiased correction is introduced within the transformation that is applicable under any test conditions.

Key words: Correlation coefficients, measurement, test characteristics, reliability, parallel forms, test equivalency.

Introduction

It has been well-established that the sample correlation coefficient, *r*, is a biased estimator of the population correlation coefficient, ρ, for normal populations, and this bias can be as much as 0.05 in absolute value under realistic research conditions (Zimmerman, Zumbo & Williams, 2003). This difference may not be vital if the research question is to simply ascertain whether a non-zero correlation exists. However, if the focus is on a precise estimate of the magnitude of a non-zero correlation in test and measurement procedures, then this discrepancy may be of concern. The Pearson product moment correlation is still commonly used as an index of reliability, exampled with parallel test forms (Coleman, 2001), test-retest conditions (Robinson-Kuropius, 2005), and inter-rater consistency (Lebreton, 2007). In such cases, calculations use a total score comprised of dichotomous or polytomous items (Kline, 2005). With increasing frequency, practitioners working in these contexts recognize sample estimates are insufficient and, therefore, are correctly utilizing the Fisher transformation to provide accompanying probabilistic inferences (Fouladi, 2002).

The motivation for this study centers on the failure of Fisher's transformation to incorporate either test length or test difficulty into confidence interval calculations. Without correction, test statistics and confidence intervals from utilizing the Fisher transformation become increasingly imprecise ultimately resulting in inflated Type I error. To date, research has neither demonstrated the inefficiencies of utilizing this method, nor further advocated a test statistic inclusive of test properties upon which to draw more accurate inferences about the population. In this article, an empirical demonstration of systemic errors between the empirical distribution and the Fisher transformation is presented which can be traced to test properties of length and difficulty. Based on the results, a correction factor inclusive of test properties is introduced and examined using a Monte Carlo simulation study to explore the performance of the corrected statistic to the existing Fisher transformation.

Jeffrey Harring is an Associate Professor in the Department of Measurement, Statistics and Evaluation. Email him at: harring@umd.edu. John Wasko is a Colonel in the U.S. Army. Email him at: john.wasko@us.army.mil.

Methodology

Pearson Correlation

The Pearson's correlation coefficient is a measure of the strength of the linear relation between two continuous variables and is defined as

$$\rho = \rho(\mathbf{x},\mathbf{y}) = \frac{Cov(\mathbf{x},\mathbf{y})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} \quad (1)$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors of scores of size $n, Cov(\mathbf{x},\mathbf{y})$ represents the population covariance and $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$ are population standard deviations. Invariably researchers report a point estimate for reliability using the form

$$\hat{\rho} = \rho(\mathbf{x},\mathbf{y}) = r = \frac{s_{\mathbf{xy}}}{s_{\mathbf{x}}s_{\mathbf{y}}},$$

where $s_{\mathbf{xy}}$, $s_{\mathbf{x}}$ and $s_{\mathbf{y}}$ are sample statistics corresponding to the population quantities in (1). For test-retest reliability let,

$$\mathbf{x} = (x_{T1},\ldots,x_{Tn}) \overset{iid}{\sim} N(\mu_{xT},\sigma^2_{xT})$$
$$\mathbf{y} = (y_{T1},\ldots,y_{Tn}) \overset{iid}{\sim} N(\mu_{yT},\sigma^2_{yT})$$

represent the total scores of n respondents administered the same test on different occasions. For parallel forms, let

$$\mathbf{x} = (x_{A1},\ldots,x_{An}) \overset{iid}{\sim} N(\mu_{xA},\sigma^2_{xA})$$
$$\mathbf{y} = (y_{B1},\ldots,y_{Bn}) \overset{iid}{\sim} N(\mu_{yB},\sigma^2_{yB})$$

represent the total scores of n respondents administered different tests on different occasions. By letting A and B represent two raters scoring the same test for n respondents would constitute inter-rater reliability. Particular to test-retest and parallel forms, it is assumed that no learning has occurred as a result of the first exam or in the interim prior to administration of the second exam.

Central Limit Theorem Application

The Pearson's correlation coefficient assumes total scores to be normally distributed; this is made possible by the central limit theorem (CLT) (see Hogg & Craig, 1995 for a full description). Reviewing its application, if $i_1, i_2, \ldots i_J$ represent the scores for a test of J items, independent and identically distributed from any distribution, then their sum

$$i_1 + i_2 + \ldots + i_J = T_0 \sim N(J\mu, J^2\sigma^2)$$

is approximately normal for sufficiently large values of J. Although sufficiently large is not a quantifiable number, this requirement is important given the need for a bivariate normal distribution upon which correlation inferences are predicated (Quereshi, 1971). A rule of thumb of J exceeding 30 items has been suggested. Not to be overlooked are the other requirements for use of the CLT. First is the requirement of independence. Conditional independence is assumed, where the likelihood a respondent answers an item correctly or incorrectly is independent of their response to any other test item. Second is the concept of identically distributed, where the collection of J items should all be dichotomously scored, $i = [0,1]$, or polytomously scored $i = [0,1,\ldots,R]$.

Even if the total score is well approximated by a normal distribution, the total score random variable is still discrete. In such cases, when making probabilistic inferences with a continuous distribution with discrete data, a continuity correction is often applied (Devore, 2000). Recall that Pearson's correlation is designed for continuous random variable pairs that follow a bivariate normal distribution. Without a sufficient number of J items, the total score distributions depart from univariate normality.

This condition is further exacerbated in extremely easy or difficult shorter tests resulting in highly skewed total scores; although this becomes less of an issue as test length increases, test difficulty affects the rate of asymptotic convergence to a normal distribution. Further, the total score variable is not continuous, it is discrete. With all statistics, when underpinning assumptions are violated, the accuracy of the results becomes increasingly questionable. Such inaccuracies are often commensurate with inflated Type I error rates. It is within this framework that the need for an item-type correction encompassing test length and difficulty and a continuity correction may be advocated.

Fisher Transformation

With

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \sim N(\mu, \Sigma),$$

following a bivariate normal distribution, define a random variable Z as

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

approximated by the following normal distribution characterized by its mean and variance

$$Z \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right),\ \frac{1}{n-3}\right).$$

Being normally distributed, these relations can be used in the traditional construction of confidence intervals and hypothesis tests. The transformation of the *r* random variable is called the Fisher transformation; the immediate discussion centers on confidence intervals, presentation of appropriate hypothesis tests are provided later.

A 2-sided $(1-\alpha)\%$ confidence interval for the true correlation, ρ, is obtained via the following steps:

1. Determine the $(1-\alpha)\%$ confidence interval for *Z* such that

   $$(1-\alpha)\%\ CI = (Z_L, Z_U)$$

   where

   $$Z_L = Z + \frac{1}{\sqrt{n-3}}\Phi^{-1}\left(\alpha/2\right)$$

   and

   $$Z_U = Z + \frac{1}{\sqrt{n-3}}\Phi^{-1}\left(1-\alpha/2\right).$$

2. Create a $(1-\alpha)\%$ confidence interval for ρ by transforming these *Z* confidence limits back onto the correlation scale

$$(1-\alpha)\%\ CI = \left(\frac{\exp(2Z_L)-1}{\exp(2Z_L)+1}, \frac{\exp(2Z_U)-1}{\exp(2Z_U)+1}\right).$$

Empirical Demonstration of Theoretical Findings

To illustrate the need to account for the number of test items for asymptotic convergence to a normal distribution, two empirical experiments are conducted. Conditions for the first simulation are a test length of $J = 25$ items, a population correlation of $\rho = 0.8$, administered to $n = 100$ respondents, where each item is an independent dichotomous response with a *p*-value of 0.60.

Conditions for the second simulation are $J = 35$, $\rho = 0.7$, $n = 100$, and a *p*-value of 0.70. For each simulation, responses for *J* items for respondent *i* ($i = 1, 2, \ldots, n$) were created according to a particular *p*-value representing a test. A second set of responses, representing a second test, were created such that each item was correlated with its first test equivalent according to a particular ρ. The item scores were totaled for each test for each respondent, resulting in a paired set of total scores of length *n*. A correlation estimate was calculated and retained for this set of total scores and, using the Fisher transform, two-sided 90% and 95% confidence intervals were calculated. Knowing the true ρ, each interval was evaluated to determine if it encompassed the true value, successes were noted. This was repeated for 10,000 trials for each experimental condition, the percentage of these successes estimates the coverage probability. Success percentages below the $(1-\alpha)\%$ specification indicates an inflated Type I error (the probability of rejecting a correct null hypothesis).

For each simulation, every sample correlation value was transformed to a *Z* random variable. A histogram of the sampling distribution is overlaid with the Fisher transform. Sampling distributions for 3rd and 4th moment statistics are provided on each plot including coverage probabilities.

Clearly, a snapshot exploring just two experimental conditions does not provide

Figure 1: Empirical *Z*-Scaled Histogram with Fisher Transform Overlay
10,000 trials, $\rho = 0.8$, n =100, test length $J = 25$, p-value = 0.6

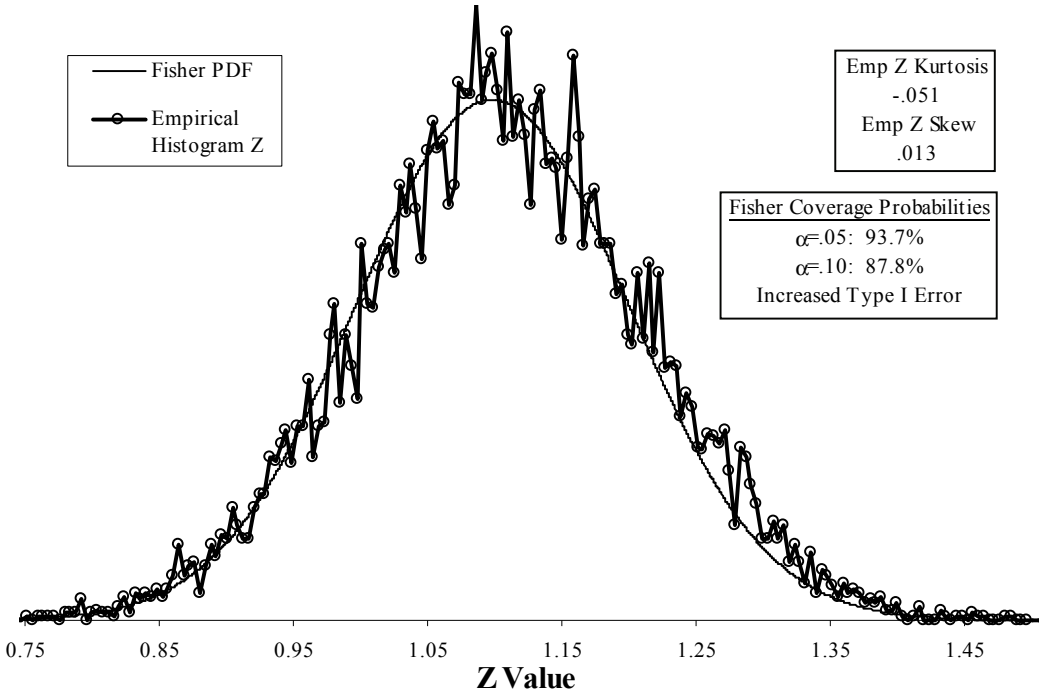
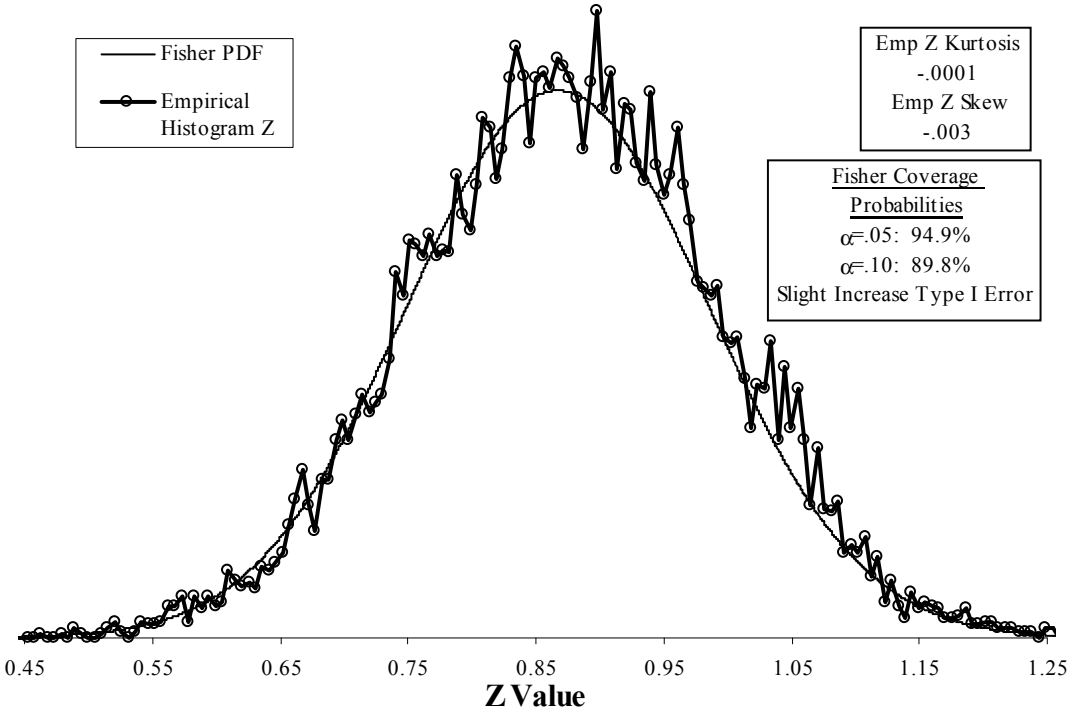


Figure 2: Empirical *Z*-Scaled Histogram with Fisher Transform Overlay
10,000 trials, $\rho = 0.7$, n =100, test length $J = 35$, p-value = 0.7



irrefutable evidence; but results highlight areas requiring further exploration.

1. The transformation of the sample correlation remains well characterized by a normal distribution.

2. There was inflated Type I error in both cases, albeit to different degrees. From these two simulations, it is difficult to tell if the results are due exclusively to sampling error, the coarseness of measurement, or a more systemic problem commensurate with the CLT requirements previously noted. Operating under the assumption the results are indicative of a systemic problem, then:

    a. It would appear that higher levels of skewness and negative kurtosis in the sampling distribution comparatively increased the Type I error. A negative kurtosis is indicative of a platykurtic distribution with larger tails. This finding is commensurate with the requirement for a sufficient number of *J* items under the CLT to subscribe to a normal distribution. Accordingly, insufficient numbers of *J* items are more likely to demonstrate skewness and kurtotic properties in the sampling distribution.

    b. In the case of very small negative kurtosis and skewness, there remains a slight inflation in Type I error. Again, assuming this is a systemic condition above and beyond sampling error, this would coincide with need for a continuity correction.

    c. There is not enough information, however, demonstrating systemic coverage probability error to suggest a parametric form for a correction or adjustment which would result in a more accurate test statistic.

To better evaluate the viability of systemic inflated Type I errors, as well as to explore a functional parametric form as a remedy, a broader, multi-factor simulation study was carried out. Retaining the finding that the Z transform of the sample correlation is reasonably represented by a normal distribution, the estimate of the μ parameter is retained. If these occurrences prove to be systemic, they can be mitigated by developing a correction to the σ parameter specified as part of the Fisher transformation.

Study Design

This multi-factor empirical study was designed to jointly assesses the performance of the Fisher transformation and explore a viable parametric form for a correction. As a result of the theoretical analysis, it was expected that the sampling statistic would be consistently negatively biased. Such a bias corresponds to an increased Type I error rate, thus substantiating the need for a continuity correction. Further, it was additionally expected that the bias would be exacerbated by some function of *J* items as *J* decreased; this would substantiate the need for an item-type correction. Subsequent steps in developing a correction would only be necessary if these expectations are observed.

Using the same factors previously noted, a wide-ranging series of experimental conditions for each factor was used. Table 1 displays the conditions under which independent dichotomous responses were generated.

Table 1: Simulation Study Experimental Conditions and Corresponding Levels

Conditions	Levels
n = number of respondents in the sample	4 levels (25, 50, 100, 200)
J = number of items on the test	4 levels (10, 20, 40, 60)
p = probability of getting the item correct	3 levels (0.50, 0.65, 0.80)
ρ = correlation between two tests	3 levels (0.60, 0.75, 0.90)

The result is $4 \times 4 \times 3 \times 3 = 144$ different experimental conditions using the same simulation process previously described. Again, 10,000 trials were conducted per condition.

As opposed to assessing probability coverage and overall sampling distribution characteristics, the differences between the sampling distribution and the Fisher transformation at various percentiles were investigated. This change was adopted for two reasons. First, the hypothesis that the Fisher transformation is inaccurate necessitates anchoring the empirical sampling distribution as the correct distribution. Second, assessment of differences at various percentiles under various treatment conditions facilitates development of a functional form for a correction. These percentiles are analogous to the most common Type I error controls in confidence interval construction and hypothesis testing, both 1-sided and 2-sided. To evaluate the distributional differences, for each set of 10,000 trials, sample correlation values were numerically ordered where

$$r_i = r_1, r_2, ...r_{10000}$$
$$r_1 \le r_2 \le r_3 . \le ..r_{10000}$$

and the following values were retained

$$(r_{100}, r_{9900}), (r_{250}, r_{9750}), (r_{500}, r_{9500}), (r_{1000}, r_{9000})$$

These are the empirical analogs to Type I error values, α, of 0.01, 0.025, 0.05, and 0.10 respectively. For each treatment condition, knowing ρ and n, corresponding r interval bounds from the Fisher transformation process were calculated corresponding to the particular α. Error was computed as

$$Error = r_{empirical,\%} - r_{Fisher,\alpha}$$

A plot of the error for all treatment conditions is provided in Figure 3. The pattern of errors, with $(1-\alpha)$ yielding positive errors and α negative errors indicates an underestimation of variance at smaller test lengths. Recognition of a pattern also provides sufficient empirical evidence of a systemic problem beyond sampling error. Although this plot shows a pattern, it is does not provide definitive relationships purely as a function of test length, failing to address test difficulty.

Basic statistic textbooks indicate that binomial distributions approximate well to a normal distribution as its expected value, *np*, exceeds some heuristic value. Using that principle, consider the expected total score or total correct as the independent variable. The expected total score is a function encompassing both test length, *J*, and test difficulty, *p*-value. For dichotomous tests,

$$\begin{aligned} \mathrm{E}(T_o) &= \sum_{i=1}^{J} p - value_i \\ &= \bar{p}J \\ &= \frac{\sum_{i=1}^{N} T_{o,i}}{N} \end{aligned}$$

For polytomous scored items, each item must follow the same scale, $r = 0, 1, 2, \ldots, R$.

$$\mathrm{E}(T_o) = \frac{\sum_{i=1}^{N} T_{o,i}}{NR}$$

A reduced number of treatment conditions using the expected total score as the independent variable are displayed in the error plot in Figure 4. Evidently, there is distinctive pattern as the expected total score decreases. This pattern is similar across all treatment conditions. Figure 5 shows another set of treatment conditions illustrating similar findings.

Dotted lines in Figure 5 indicate bias as a result of failure to implement a continuity correction. This correction remains constant regardless of the E(T_o) value. Additionally, there is a systemic increase in error as the expected total number of correct items decreases. This decaying relationship asymptotes to the continuity correction value as E(T_o) increases. These empirical results reinforce the theoretical findings noted when data deviate from required conditions in applying the CLT. Because these graphs are presented as a separate set of

Figure 3: Error versus Test Length across All Treatment Conditions

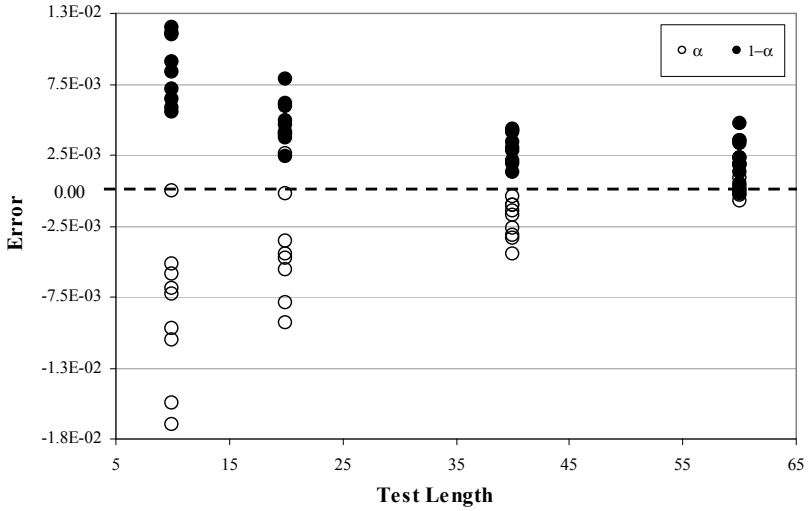


Figure 4: Error versus Expected Total Score across a Reduced Number of Experimental Conditions

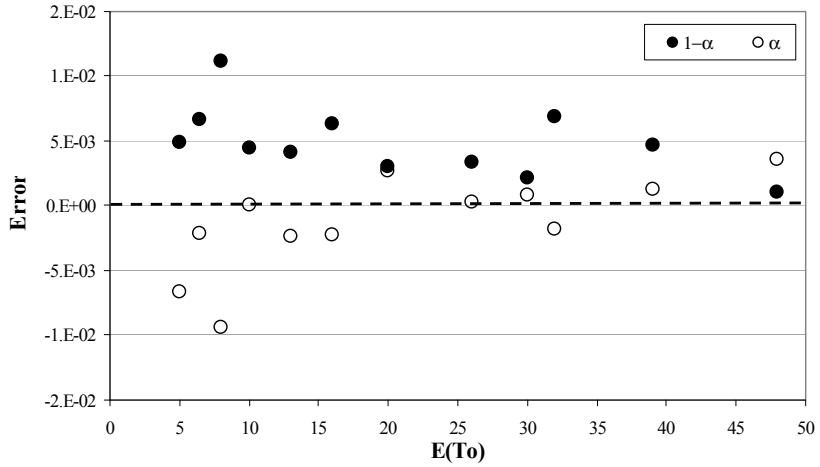
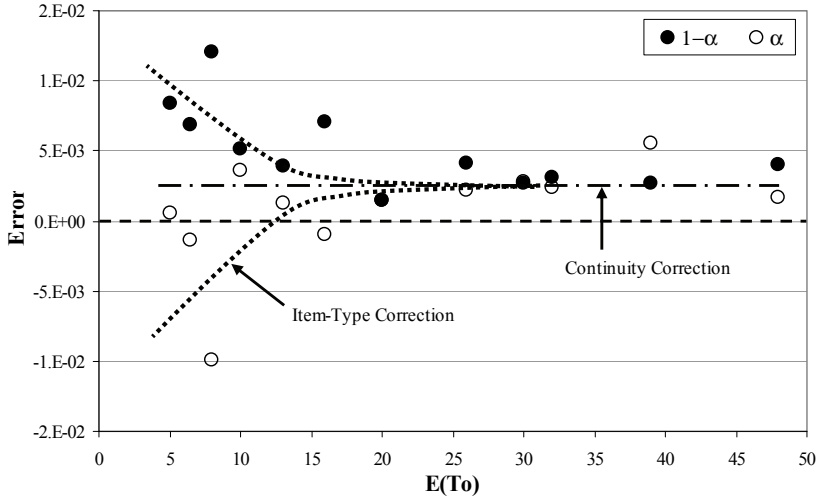


Figure 5: Error versus Expected Total Score Indicating Parametric Corrections



snapshots, there is a third relation observed which cannot be easily illustrated. Although each plot consistently exhibits a decaying relationship as E(T_o) increased, the amount and rate of decay differed conditioned upon the *p*-value or test difficulty treatment conditions. Higher *p*-values exhibited greater errors at lower E(T_o) values and took slightly longer to converge to the continuity correction. These findings are consistent with previous CLT discussions.

Proposed Correction

Though illustrating the need for a correction when applying Fisher's transformation inclusive of test properties is informative, its value is only realized with a corresponding remedy. Thus, the distributional properties of the Fisher transformation with independence of its first two moments are maintained. The item-type correction and continuity correction are independent corrections and can be treated as such in a specified solution. The impact of the *p*-value on the rate of change only affects the item-type correction. Accordingly, Fisher's transform is retained as

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

but, as opposed to utilizing the form $\sigma_Z = \frac{1}{\sqrt{n-3}}$, a corrected form is derived as

$$\sigma_Z^* = \left(\frac{\ln\left(\left(\frac{1}{1+a(pval-.5)^2} * bE(T_o)\right)+1\right)}{\ln\left(\frac{1}{1+a(pval-.5)^2} * bE(T_o)\right)} + c\right)\left(\frac{1}{\sqrt{n-3}}\right)$$

where *a*, *b*, and *c* are undetermined constants. The *a* term is associated with the *p*-value's effect on the amount and rate of decay associated with E(T_o). The *b* term is associated with the general rate of decay as the item-type or E(T_o) correction. The *c* term is associated with the continuity correction. Note that the overall correction

limit

$$_{E(To)\to\infty}\left(\frac{\ln\left(\left(\frac{1}{1+a(pval-.5)^2} * bE(T_o)\right)+1\right)}{\ln\left(\frac{1}{1+a(pval-.5)^2} * bE(T_o)\right)} + c\right) = 1 + c$$

is commensurate with the error plots previously presented. More specifically, the term

$$\frac{\ln\left(bE(T_o)+1\right)}{\ln\left(bE(T_o)\right)}$$

represents the decaying relation associated with E(T_o). Because these relations change as a function of the *p*-value, the following is introduced within the logarithm

$$\frac{1}{1+a(pval-.5)^2}$$

Figure 8 displays the correction factor shown for differing *p*-values.

Although the effect on the rate of decay is symmetrical around 0.50, the overall correction is not due to the effect of the *p*-value in the E(T_o) calculation. Figure 9 illustrates this lack of symmetry for 3 different tests lengths under a range of average *p*-values.

Other parametric representations may also be available for the correction. This choice appeared reasonable and parsimonious based on the observations of the errors between the empirical distributions and an uncorrected Fisher transform. Values for these constants were determined via an iterative process minimizing the total squared error across all treatment conditions of the form.

$$Total\ Error = \sum_{n=1}^{4}\sum_{l=1}^{4}\sum_{k=1}^{3}\sum_{j=1}^{4}\sum_{i=1}^{8}\left(r_{empirical,\%,,ijkln} - r_{Fisher^*,ijkln}\right)^2 \quad (3)$$

where *i* corresponds to the values of α, *j* represents the test length, *k* denotes the *p*-values

Figure 8: Z Standard Deviation Correction versus Number of Correct Items for Various *p*-values

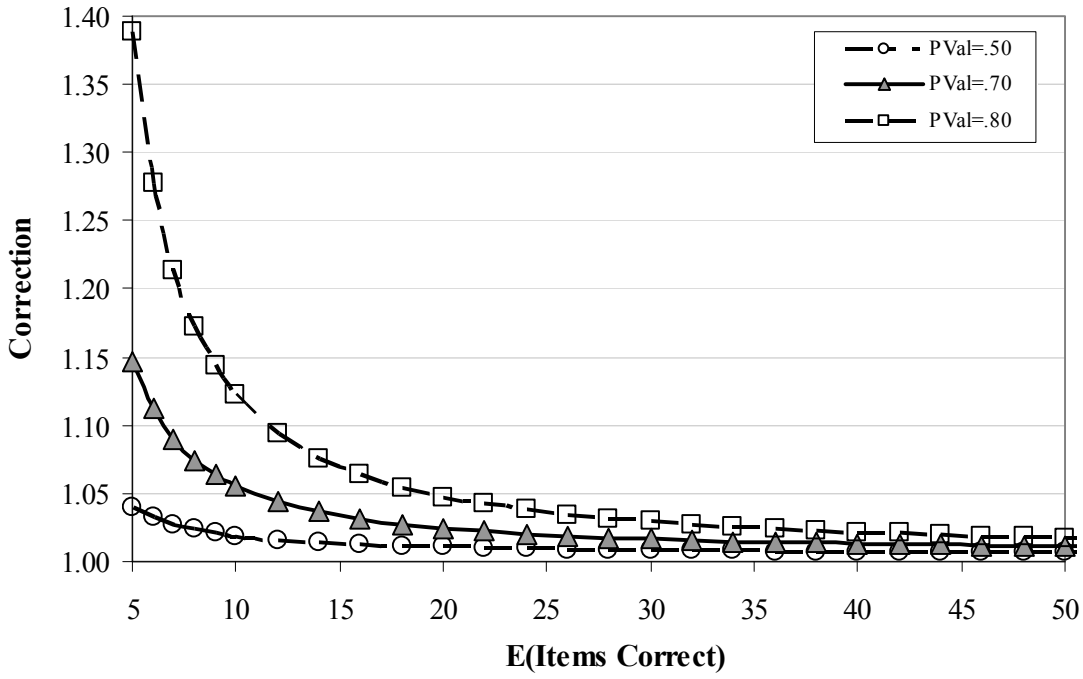
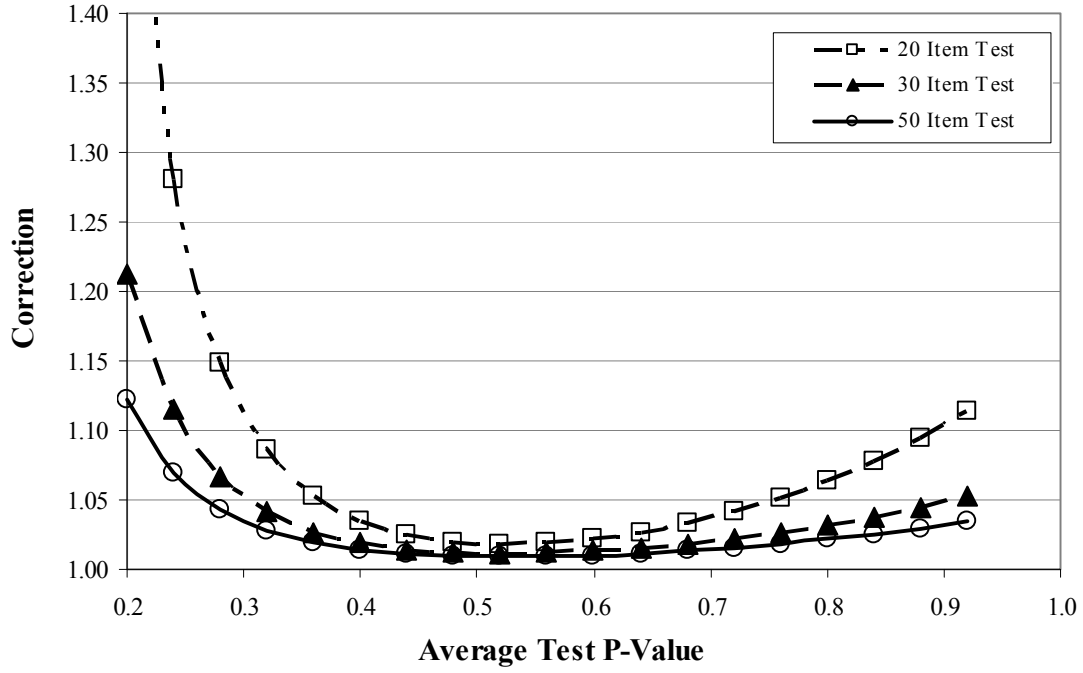


Figure 9: Z Standard Deviation Correction versus *p*-values for Various Test Lengths



for each test item, l represents the true correlation between items on each test, and n denotes the number of examinees. An evolutionary solver add-in to Excel from Frontline systems was utilized searching within a range of acceptable values. This particular solver is well suited to handle this nonlinear, mixed integer optimization problem. The resulting minimized error solution takes the form of

$$\sigma_Z^* = \left(\frac{\ln\left(\left(\frac{1}{1+40(pval-.5)^2} * 2.25E(T_o)\right)+1\right)}{\ln\left(\frac{1}{1+40(pval-.5)^2} * 2.25E(T_o)\right)} + .005 \right)\left(\frac{1}{\sqrt{n-3}}\right) \quad (4)$$

Results

Correct Assessment

Although the strategy in advocating a parametric correction is valid, it suffers from two flaws. First, the constants selected were optimized based on a set of 144 treatment conditions. As a means of cross-validation, this correction should be assessed under a different set of treatment conditions. Second, and more importantly, is the aspect of coverage probability. Reduced distributional errors resulting from an adjusted standard deviation in the Z transform does not necessarily correspond to a definitive improvement in coverage probability.

By utilizing aspects of both previous simulations, both flaws are addressed and a more thorough assessment of the proposed correction is provided. Using the same factors, consider next a broader series of treatments for each factor. Independent dichotomous responses were generated under the following conditions enumerated in Table 2.

The result is $5 \times 4 \times 4 \times 3 = 240$ different treatment conditions using the same process. Using both the Fisher transform and the proposed correction, two-sided 90%, 95%, and 99% confidence intervals were calculated from the sample correlation value used in this study. Knowing the true ρ for each trial an assessment was made as to whether this value was within the Fisher and the corrected interval, noting successes. This was repeated for 10,000 trials for each simulation resulting in an estimate of the coverage probability. Success percentages below the $(1-\alpha)\%$ specification indicate an inflated Type I error.

As formal statistical assessments of these coverage probabilities, performance in terms of bias and mean square error across all conditions was considered. Bias is defined as $Bias(\hat{\theta},\theta) = E(\hat{\theta}-\theta)$, where θ is the specified confidence interval, 99%, 95% or 90%, and $\hat{\theta}$ represents the proportion of intervals containing the true population correlation value separately for the Fisher transformation and the proposed correction.

Mean square error (MSE) is determined by: $MSE = V(\hat{\theta}) + Bias^2$ where $V(\hat{\theta})$ is the variance of the estimates determined across the set of the treatment conditions.

Graphical summaries in Figures 10a, 10b, and 10c are presented as boxplots of coverage probability results from the conditions over each of the 3 test related parameters associated in calculating the proposed formula: sample size of respondents (n), expected number of items correct ($E(T_o)$), and an average test p-value, respectively.

Table 2: Simulation Study Experimental Conditions and Corresponding Levels

Conditions	Levels
n = number of respondents in the sample	5 levels (25, 50, 100, 200, 400)
J = number of items on the test	4 levels (10, 20, 40, 80)
p = probability of getting the item correct	3 levels (0.50, 0.60, 0.70, 0.80)
ρ = correlation between two tests	3 levels (0.65, 0.75, 0.85)

Figure 10a: Side-by-Side Boxplots of Coverage Probability Error Comparison at $\alpha = 0.01$ Over Expected Correct Items across All Conditions

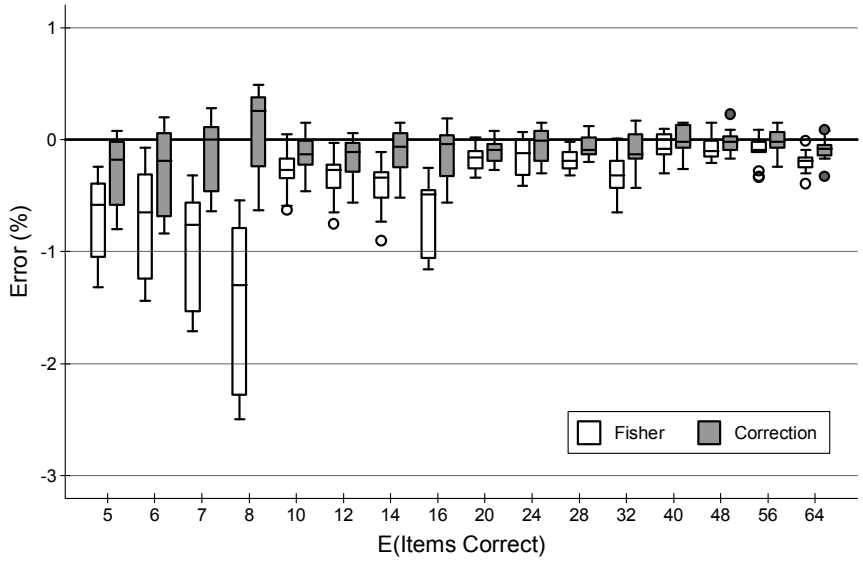
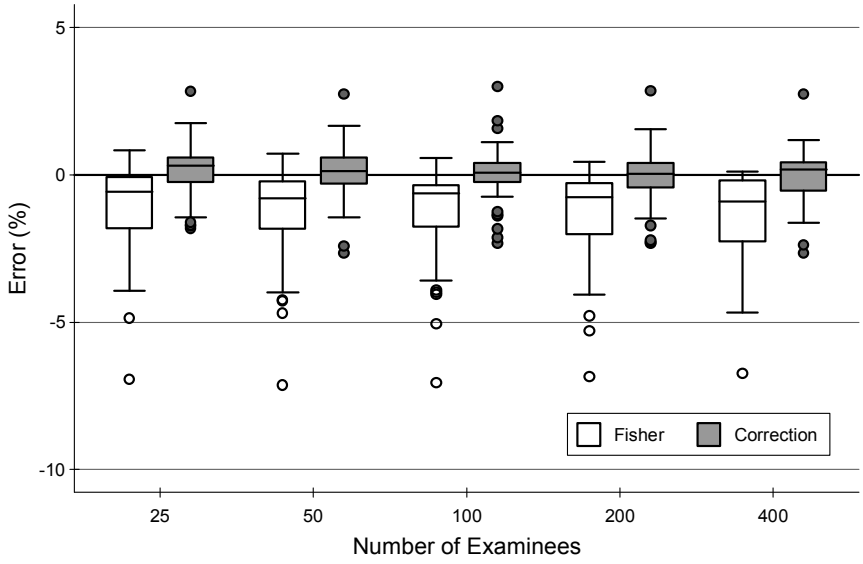


Figure 10b: Side-by-Side Boxplots of Coverage Probability Error Comparison at $\alpha = 0.05$ over average *p*-value across All Conditions



Summary results are shown in Table 3, with bias and mean squared error values provided across all conditions. The results showed improvement over the uncorrected Fisher transformation with 10 times less bias and a total reduction of error exceeding 500% across all conditions. These improvements are also consistent with each of the 28 cross-classified results, outperforming the Fisher transform with smaller bias and mean square error.

Table 3: Bias and MSE for Fisher's Transformation and the Proposed Correction for All Experimental Conditions

Description	Fisher Transformation		Proposed Model	
	Bias	MSE	Bias	MSE
Overall	-0.936	2.285	-0.095	0.443
By Sample Size				
25	-0.887	2.168	-0.060	0.447
50	-0.929	2.267	-0.089	0.451
100	-0.937	2.261	-0.098	0.435
200	-0.965	2.422	-0.120	0.469
400	-0.960	2.348	-0.107	0.422
By *p*-value				
0.50	-0.658	1.206	-0.214	0.432
0.60	-0.739	1.403	-0.223	0.379
0.70	-0.916	2.060	-0.051	0.352
0.80	-1.431	4.494	0.109	0.614
By Alpha				
0.01	-0.423	0.396	-0.096	0.061
0.05	-1.105	2.471	-0.195	0.427
0.10	-1.279	3.999	0.007	0.844
By $E(T_0)$				
5	-1.535	3.605	-0.574	1.123
6	-1.730	4.358	-0.587	1.012
7	-2.116	6.464	-0.082	0.905
8	-3.115	13.403	0.667	1.714
10	-0.703	1.018	-0.276	0.495
12	-0.779	0.963	-0.285	0.327
14	-0.915	1.332	-0.110	0.323
16	-1.612	3.766	-0.151	0.530
20	-0.294	0.148	-0.066	0.060
24	-0.314	0.228	-0.052	0.110
28	-0.420	0.353	-0.030	0.129
32	-0.696	0.702	-0.087	0.154
40	-0.098	0.077	0.060	0.072
48	-0.133	0.086	0.033	0.083
56	-0.214	0.164	0.006	0.091
64	-0.300	0.164	0.006	0.091

Though the proposed correction is empirically unbiased, it cannot be theoretically demonstrated as an unbiased estimator. Given the variety of treatment conditions examined, a theoretical proof becomes difficult without many simplifying assumptions. Some additional comments regarding a theoretical assessment include:

1. Although the need for correction based on the expected total number of items correct and the average *p*-value of the testing instrument has been theoretically and empirically demonstrated, a proper parametric form to implement such correction into probability coverage is not clear. As noted previously, there are other parametric forms which may be considered. Also, recall that the assumption of normality upon transform is still operating, which becomes more tenuous in low number of test items and extreme *p*-values. Other distributional forms can be considered upon which one would make probabilistic inferences. Finally, regarding parametric forms and distributions, this discussion is predicated that there exists a common distribution characterized by respondents and test conditions which results in an unbiased, consistent estimator controlling Type I error.

2. Due to confidence the Fisher transformation is incomplete without inclusion of summary test information in its calculations, the empirical distribution of the sample correlation values were treated as the true distribution. This was also necessary to assess systemic errors in the development of a functional parametric form for a correction. This reference empirical distribution has sampling error, which has been minimized given the large number of trials.

3. Estimates via a complex evolutionary search method were obtained from the Frontline Premium Solver add-in for the Excel Solver. Determining a so-called best set of parameter estimates for a complex nonlinear optimization required parameter constraints and other considerations in order to achieve convergence.

Based on these findings, when reporting sample Pearson product moment correlations for dichotomous and polytomously scored items, the adjustment in (4) is recommended; it is well characterized by a normal distribution. These corrections provide robust results due to violations in the application of the central limit theorem. It further provides a researcher inclusion of summary test information into any inferential statistics. Unfortunately, because of the transformation process, simple reporting of the standard error is uninformative. As such, presented below are two examples which should be used as the proper mechanism for reporting sample correlation properties.

Applications: Parallel Test Forms

Forms A and B of a particular test are each administered to 70 respondents from the same population. Each test consists of 25 items and both test are polytomously scored on a scale of [0, 1, …, 4]. The average score for form A was 41 and 45 for form B. The sample correlation was r = 0.82, and it is desired to report a 95% confidence interval for the population correlation. Z is computed with accompanying standard deviation:

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+.82}{1-.82}\right) = 1.157$$

$$\sigma_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{70-3}} = .1222$$

Next, the proposed correction is determined, which takes the form

$$\left(\frac{\ln\left(\left(\frac{1}{1+40(0.43-.5)^2}\cdot 2.25\cdot 10.75\right)+1\right)}{\ln\left(\frac{1}{1+40(0.43-.5)^2}\cdot 2.25\cdot 10.75\right)}+.005\right) = 1.016$$

where

$$E(T_o) = \frac{41+45}{(2)(4)} = 10.75$$

and

$$pval = .5*\left(\frac{41}{100}+\frac{45}{100}\right) = 0.43,$$

therefore the estimate for the standard deviation of the transformation becomes:

$$\sigma_Z^* = 0.1222*1.016 = 0.1242.$$

Because Z follows a normal distribution, a traditional 95% confidence interval for Z can be computed as follows

$$\begin{aligned} Z_L^* &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + .1242*\Phi^{-1}\left(\alpha/2\right) \\ &= 1.157 + .1242(-1.96) = .9136 \end{aligned}$$

$$\begin{aligned} Z_U^* &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + .12441*\Phi^{-1}\left(1-\alpha/2\right) \\ &= 1.157 + .12441(1.96) = 1.40 \end{aligned}$$

which can be back transformed into intervals for the population correlation

$$\begin{aligned} &(1-\alpha)\%\ CI\ for\ \rho = \\ &= \left(\frac{\exp(2Z_L^*)-1}{\exp(2Z_L^*)+1}, \frac{\exp(2Z_U^*)-1}{\exp(2Z_U^*)+1}\right) \\ &= \left(\frac{\exp(2*.9136)-1}{\exp(2*.9136)+1}, \frac{\exp(2*1.40)-1}{\exp(2*1.40)+1}\right) \\ &= (0.723, 0.886). \end{aligned}$$

The uncorrected confidence interval is $(1-\alpha)\%\ CI\ for\ \rho = (0.725, 0.885)$. The reporting should include both the sample correlation estimate and the corresponding interval values.

Applications: Inter-rater Reliability

Suppose two graders score an exam consisting of 20 dichotomous items administered to 125 respondents. The average score for each grader was 17 and the sample correlation was $r = 0.77$. Test the hypothesis the population correlation between the two graders exceeds the minimally desired reliability value of at least 0.70 at significance level of 0.05.

Using a similar process to determine the standard deviation for the proposed correction, the Fisher transformation of the standard deviation is

$$\sigma_z = \frac{1}{\sqrt{125-3}} = \frac{1}{\sqrt{125-3}} = .0905.$$

The corrected standard deviation is

$$\left(\frac{\ln\left(\left(\frac{1}{1+40(0.85-.5)^2}\cdot 2.25\cdot 16.5\right)+1\right)}{\ln\left(\frac{1}{1+40(0.85-.5)^2}\cdot 2.25\cdot 16.5\right)} + .005\right) = 1.08$$

where

$$E(T_o) = 16.5$$

and

$$pval = \left(\frac{17}{20}\right) = .85.$$

Therefore, the estimate for the corrected standard deviation of the transformation becomes

$$\sigma_Z^* = .0905*1.08 = .0978$$

and Z^* is determined via

$$\begin{aligned} Z^* &= \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_o}{1-\rho_o}\right)}{.0978} \\ &= \frac{\frac{1}{2}\ln\left(\frac{1+.77}{1-.77}\right) - \frac{1}{2}\ln\left(\frac{1+.70}{1-.70}\right)}{.0978} \\ &= \frac{1.0203-.8673}{.0978} = 1.564. \end{aligned}$$

Because

$$Z^* \le Z_{crit,1-\alpha}$$

$$1.564 \le 1.644$$

the null hypothesis H_o is retained. It appears these graders do not meet the minimally acceptable inter-rater reliability. Corrective actions, such as additional grader training,

would be required in such cases. However, the hypothesis test without the correction results in

$$Z^* \geq Z_{crit,1-\alpha}$$
$$1.691 \geq 1.644.$$

In contrast to the results using the correction, the null hypothesis would be incorrectly rejected. Multiple rater comparisons or multiple parallel forms may as well be addressed with this correction using a multiple comparison Type I error adjustment such as Bonferroni or Tukey.

Because the proposed correction occurs within the Z transform (see Figures 8 and 9), it is difficult to interpret its impact in the original correlation scale. The width of a correlation confidence interval is not only a function of *r*, α, and *n*, but this study has demonstrated $E(T_o)$ and the average *p*-value as well. To better understand the effects of this correction in the desired scale, the following 3D plots show the difference in CI widths between the Fisher transformation and this correction, where the proposed correction always result in larger widths in order to maintain an accurate Type I error control. In each plot, *r* was 0.75 and α was 0.05. The range of test items used coincides with test section lengths of the major standardized educational exams such as the SAT, GRE, LSAT, and MCAT.

Conclusion

The Fisher transformation is remarkably efficient, yet was not designed with an intended use of summed dichotomous or polytomous data. This correction accounts for departures from asymptotic convergence under the central limit theorem due to test length and average item difficulty. Further, this correction can be easily applied, providing substantially more accurate results over the Fisher transformation. This study also illustrates the coarseness of dichotomous measures has no effect on the coverage probability results of the true population correlation as this is accounted for in the correction and results from application of the central limit theorem.

For those positing a unidimensional construct, the use of Pearson correlation can be easily extended to allow for items which load differently on the latent dimension. By weighting each item and making an adjustment to the total score, an omnibus reliability measure based on total score can be obtained.

Throughout the study, a homogeneous *p*-value for each test item was used. Because most tests are comprised of items with varying *p*-values, the performance of this correction was examined under a wide range of *p*-value distributions. This robust analysis explored extreme deviations from the simulation conditions, using a highly kurtotic uniform distribution and bi-modal distributions with different expected average *p*-values. The results for this analysis are present in Appendix A and reaffirm the use of this correction under any conditions.

Though the proposed correction is easily implemented with demonstrated efficiency across a wide range of test conditions, a nonparametric alternative is also available. Nonparametric bootstrap methods remain a viable option for researchers desiring confidence interval estimates; whereas such options might also produce robust results, they require both sufficient data and custom coding.

References

Barnette, J. J. (2005). ScoreRel CI: An Excel program for computing confidence intervals for commonly used score reliability coefficients. *Educational and Psychological Measurement*, *65*, 980-983.

Colman, A. M. (2001). *A dictionary of psychology*. Oxford University Press: Great Britain.

Denton, G., Durning, S., & Hemmer, P. (2004). A call for use of confidence intervals with correlation coefficients. *Teaching and Learning in Medicine*, *16*, 111-112.

Devore, J. L. (2000). *Probability and statistics for engineering and the sciences* (*5th Ed.*). Pacific Grove, CA: Duxbury.

Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes. *Educational and Psychological Measurement*, *61*, 517-531.

Figure 13: Confidence Interval Width Difference between Proposed Correction and Fisher Transform at $n = 150$

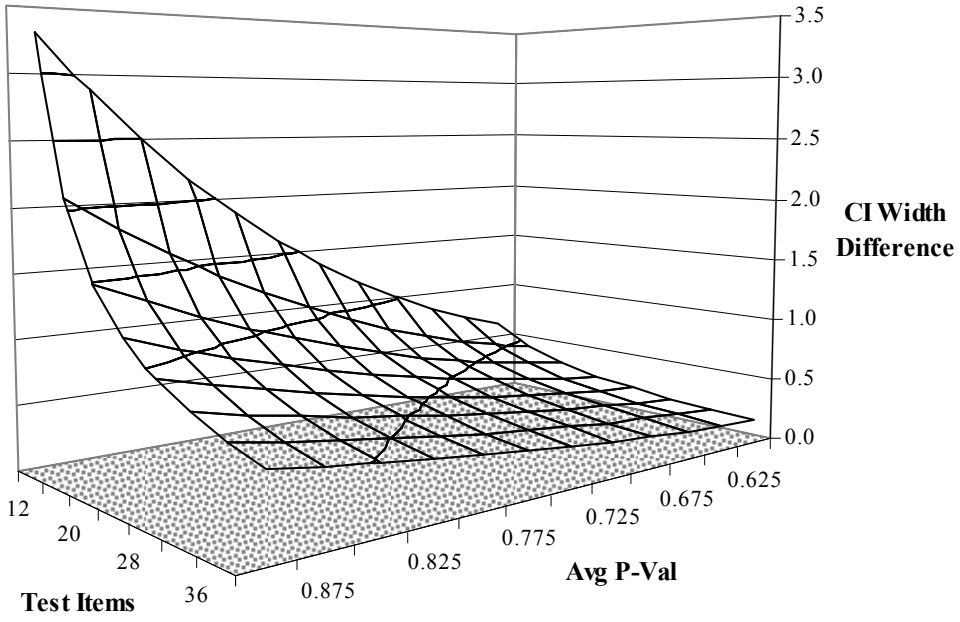
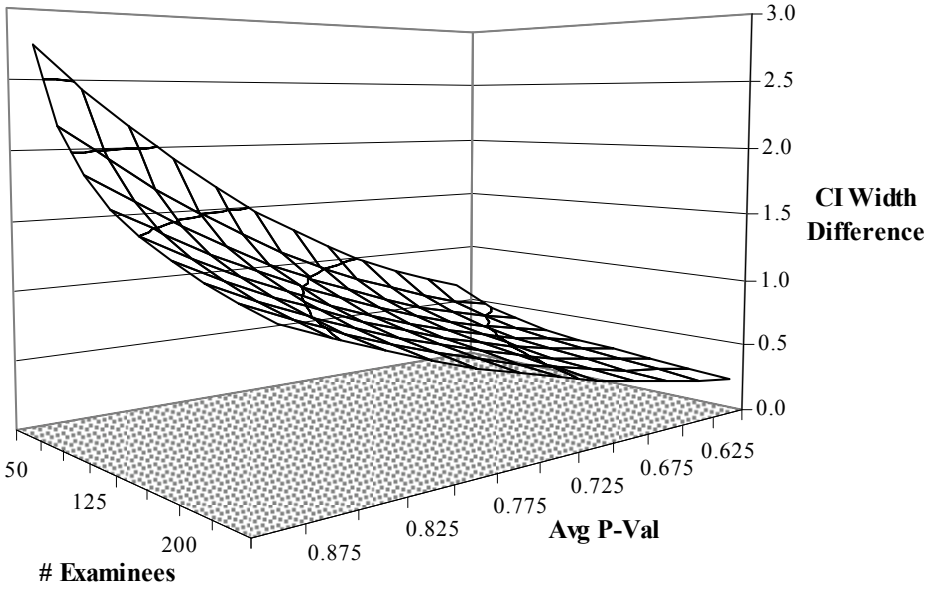


Figure 14: Confidence Interval Width Difference between Proposed Correction and Fisher Transform at $J = 20$



Fouladi, R. T., Marani, S. K., & Steiger, J. H. (2002). Moments of the Fisher transform: applications using small samples. *American Statistical Association Proceedings of the Joint Statistical Meetings [CD-ROM]*, 1032-1037.

Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics* (*5th Ed*.). Englewood Cliffs, NJ: Prentice Hall.

Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Laubscher, N. F. (1959). Note on Fisher's transformation of the correlation coefficient. *Journal of the Royal Statistical Society, Series B (Methodological*), *21*, 409-410.

Lebreton, J. M., & Senter, J. (2007). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods Online First*, 1-38.

Mulry, M. H., & Wolter, K. M. (1981). The effect of Fisher's Z transformation on confidence intervals for the correlation coefficient. *U.S. Bureau of the Census*, 601-608.

Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, *29*, 213-225.

Quereshi, M. Y. (1971). Note on the Pearson r as a function of the bivariate distributional characteristic. *Journal of Educational Measurement*, *8*(*3*), 142-147.

Robinson-Kurpius, S. E., & Stafford, M. E. (2005). *Testing and measurement: A user-friendly guide*. Thousand Oaks, CA: Sage.

Shen, D., & Lu, Z. (2005). Computation of correlation coefficient and its confidence interval in SAS. *SAS Paper*, *170-31*. SAS Institute.

Task Force on Reporting of Research Methods in AERA Publications (2006). *Standards for reporting on empirical social science research in AERA publications*. Washington, DC: American Educational Research Association.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25-32.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications, Volume 3*. Thousand Oaks, CA: SAGE.

Wilkinson, L. (1999). The American Psychological Association Task Force on Statistical Inference, Statistical Methods in Psychology: Guidelines and Explanations. *American Psychologist*, *54*, 594-604.

Winterbottom, A. (1979). A note on the derivation of Fisher's transformation of the correlation coefficient. *The American Statistician*, *33*, 142-143.

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica*, *24*, 133-158.

Appendix

As a means of robust analysis, the proposed correction was explored under 4 different sets of varied p-values. Empirical treatments remained unchanged for sample size, population correlation, and test length. However, instead of a homogeneous *p*-value for each item on a test of length *J*, the following were considered:

a. *p-value* = 0.50 per test item to a bimodal distribution of the following form

$$\frac{J}{2}Unif(.2-.4)+\frac{J}{2}Unif(.6-.8)$$

per test. *P*-values were redrawn from this distribution for each trial. The average p-value is 0.50.

b. *p-value* = .60 per item to a distribution of the form

$$Unif(.3-.9)$$

per test, redrawn for each trial. The average *p-value* is 0.60.

c. *p-value* = 0.70 per item to a distribution of the form

$$\frac{J}{2}Unif(.45-.65)+\frac{J}{2}Unif(.75-.95)$$

per test, redrawn for each trial. The average *p*-value is 0.70.

d. *p-value* = 0.80 per item to a distribution of the form

$$Unif\,(.65-.95)$$

per test, redrawn for each trial. The average *p-value* is 0.80.

Collective results are presented in the Table 4. Similar to this validation study, in bias and mean square error, overall and across each of treatment conditions, the proposed correction outperformed the Fisher transformation. Further, the Type I error of the Fisher transformation is comparatively higher compared with a test of items with homogeneous *p*-values. This reaffirms the suitability of this correction under any conditions, regardless of the *p*-value distribution underpinning the test items.

Table 4: Robust Analysis for Extreme *p*-values; Bias and MSE for Fisher's Transformation and the Proposed Model across All Experimental Conditions

Description	Fisher Transformation		Proposed Model	
	Bias	MSE	Bias	MSE
Overall	-1.100	3.081	-0.229	0.698
By Sample Size				
25	-1.020	2.615	-0.169	0.574
50	-1.143	3.231	-0.260	0.727
100	-1.078	3.031	-0.204	0.694
200	-1.097	3.156	-0.220	0.678
400	-1.164	3.423	-0.291	0.837
By P-value				
0.50	-0.929	2.125	-0.461	0.837
0.60	-0.896	2.007	-0.341	0.636
0.70	-1.086	3.005	-0.191	0.612
0.80	-1.490	5.217	0.078	0.718
By Alpha				
0.01	-0.522	0.578	-0.166	0.114
0.05	-1.253	3.284	-0.318	0.721
0.10	-1.526	5.395	-0.202	1.266
By $E(T_0)$				
5	-2.116	6.560	-1.087	2.448
6	-2.027	6.026	-0.760	1.669
7	-2.495	9.539	-0.371	1.786
8	-3.451	16.079	0.493	1.816
10	-1.037	1.530	-0.591	0.659
12	-1.001	1.667	-0.480	0.722
14	-1.163	2.013	-0.331	0.490
16	-1.620	3.918	-0.182	0.744
20	-0.422	0.379	-0.174	0.216
24	-0.427	0.310	-0.163	0.122
28	-0.508	0.428	-0.098	0.133
32	-0.683	0.769	-0.073	0.239
40	-0.140	0.063	0.008	0.052
48	-0.131	0.058	0.040	0.058
56	-0.178	0.095	0.035	0.077
64	-0.244	0.172	0.074	0.116