Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 16

5-1-2007

A Comparison of Eight Shrinkage Formulas under Extreme Conditions

David A. Walker
Northern Illinois University, dawalker@niu.edu

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 162-172

1538 – 9472/05/$95.00

A Comparison of Eight Shrinkage Formulas under Extreme Conditions

David A. Walker
Northern Illinois University

The performance of various shrinkage formulas for estimating the population squared multiple correlation coefficient (ρ^2) were compared under extreme conditions often found in educational research with small sample sizes of 10, 15, 20, 25, 30 and regressor variates ranging from 2 to 4. A new formula for estimating ρ^2, Adj R^2_{DW}, was examined in terms of its performance under various conditions of N, p, ρ^2, along with its bias properties and standard error estimates. The two shrinkage formulas that performed most consistently were the Claudy (Adj R^2_C) and Walker (Adj R^2_{DW}).

Key Words: Adjusted R^2, shrinkage, population squared multiple correlation

Introduction

Various shrinkage formulas for estimating the population squared multiple correlation coefficient (ρ^2) has been the topic of interest (cf. Carter, 1979; Claudy, 1978; Huberty & Mourad, 1980; Lucke & Embretson, 1984). The purpose of this article is to compare the performance of eight shrinkage formulas for estimating the population multiple correlation coefficient with small sample sizes of 10, 15, 20, 25, 30 and with regressor variates ranging from 2 to 4. Small sample sizes were used because in applied research fields, such as educational research, these sample conditions often are encountered (Claudy, 1972; Huberty & Mourad, 1980). Also, regressor variates were chosen to be between 2 and 4 for the same reason cited formerly with sample size; typicality of conditions frequently encountered in educational research.

David Walker is Associate Professor of Educational Research and Assessment at Northern Illinois University. His research interests include structural equation modeling, effect sizes, factor analyses, predictive discriminant analysis, predictive validity, weighting, and bootstrapping. Email him at P60DAW1@wpo.cso.niu.edu

The sample squared multiple correlation coefficient, or R^2, indicates the percentage of variance in the dependent variable explained by the linear combination of the independent variables. R^2 has been found to overestimate the population multiple correlation (ρ^2) and, hence, is seen as an upwardly biased approximation of ρ^2 with limited accuracy (Agresti & Finlay, 1997; Pedhazur, 1997). This overestimation has been linked to the problem of error, often either measurement or sampling error, connected to the variability found in random independent variables (Claudy, 1972), related to sample size, and associated with the number of X variables in a model (Huberty & Mourad, 1980; Shumacker, Mount, & Monahan, 2002). The population multiple correlation can be expressed as (Browne, 1975):

$$\rho^2 = \text{corr}^2\{Y, \sim Y(X|ß_0, ß)\} \qquad (1)$$

where,

Y = Dependent variable
X = Set of regressors
ß = Population regression weights

Due to amending for this overestimation, the adjusted R^2 (adj R^2) has been used as a more accurate method than R^2 for estimating ρ^2. That is, the adj R^2 is more exact than R^2 due to its correction for shrinkage and its ability to produce an accurate estimate of the population value for ρ^2. Adjusted R^2 can be expressed as (Agresti & Finlay, 1997):

$$R^2_{adj} = R^2 - \frac{p-1}{N-p} * (1 - R^2) \qquad (2)$$

Other shrinkage formulas for estimating the population multiple correlation coefficient have been presented with the goal of reducing the positive bias of R^2. As noted by Carter (1979), many of the subsequent formulas are decidedly related algebraically and/or are hybrids of one another.

Formulas 3 to 6 and 9 are reproduced in Huberty and Mourad (1980). According to Huberty and Mourad, Smith proposed, but presented by Ezekiel (1929), the first adjusted R^2 shrinkage formula, R^2_S, where:

$$R^2_S = 1 - \frac{N}{N-p-1} * (1 - R^2) \qquad (3)$$

Ezekiel (1930) proposed R^2_E, where:

$$R^2_E = 1 - \frac{N-1}{N-p-1} * (1 - R^2) \qquad (4)$$

Wherry (1931) proposed R^2_W, where:

$$R^2_W = 1 - \frac{N-1}{N-p} * (1 - R^2) \qquad (5)$$

Olkin and Pratt (1958) proposed R^2_{OP}, where:

$$R^2_{OP} = 1 - \frac{N-3}{N-p-1} * (1 - R^2) - \frac{2(N-3)}{(N-p-1)(N-p+1)} * (1 - R^2)^2 \qquad (6)$$

Pratt (1964 as cited in Claudy, 1978) proposed R^2_P, where:

$$R^2_P = 1 - \frac{(N-3) * (1 - R^2)}{N-p-1} * 1 + \frac{2(1 - R^2)}{(N-p-2.3)} \qquad (7)$$

Herzberg (1969 as cited in Claudy, 1978) proposed R^2_H, where:

$$R^2_H = 1 - \frac{(N-3) * (1 - R^2)}{N-p-1} * 1 + \frac{2(1 - R^2)}{(N-p+1)} \qquad (8)$$

Claudy (1978) proposed R^2_C, where:

$$R^2_C = 1 - \frac{N-4}{N-p-1} * (1 - R^2) - \frac{2(N-4)}{(N-p-1)(N-p+1)} * (1 - R^2)^2 \qquad (9)$$

Walker (2006) proposed R^2_{DW}**,** which is an algebraic alteration of R^2_C and, hence, N - 4.15 was a more optimal empirical modification of N – 4 than N - 5**,** where:

$$R^2_{DW} = 1 - \frac{N-4.15}{N-p-1} * (1 - R^2) - \frac{2(N-4.15)}{(N-p-1)(N-p+1)} * (1 - R)^2 \qquad (10)$$

where,
N = Sample size
p = Number of X variables
R^2 = Multiple correlation coefficient

Methodology

Via a simulation program written in SPSS (Statistical Package for the Social Sciences) v. 12.0, the following study reviewed the shrinkage performance of the eight multiple correlation estimators noted previously when ρ^2 is known at .15, .30, .45, .60, .75, .90, N = 10, 15, 20, 25, 30, p = 2, 3, 4, under normal distributional assumptions, and where the number of iterations within the simulation was 500.

Results

Overall, the study's findings indicated that all of the eight shrinkage formulas utilized under the research's specified conditions did succumb to bias, as was expected, either via under or overestimation of the population multiple correlation. Table 1 indicates that the two most consistently accurate formulas were Claudy and Walker. When looking at small sample sizes with few predictors with a $\rho^2 \leq .45$, Table 1 shows that the Smith, Ezekiel, Wherry, and Olkin and Pratt formulas typically underestimated, often times greatly, ρ^2 in comparison to the Pratt, Herzberg, Claudy, and Walker formulas. However, the Pratt and Herzberg formulas tended to overestimate the population multiple correlation at .60, .75., and .90, respectively, regardless of the sample size and especially when p = 2 and 3. The Claudy and Walker formulas were consistently accurate in these same conditions, with only a small portion of overestimation when p = 2.

Table 1. Values for Eight Shrinkage Formulas when N = 10 to 30, p = 2 to 4

N = 10, p = 2

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.214	-.093	.044	-.011	.134	.199	.181	.155
.300	.000	.100	.213	.191	.307	.389	.357	.324
.450	.214	.293	.381	.383	.471	.572	.528	.484
.600	.429	.486	.550	.564	.627	.747	.693	.636
.750	.643	.679	.719	.736	.774	.914	.854	.779
.900	.857	.871	.888	.898	.912	1.000	1.000	.915

N = 15, p = 2

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.063	.008	.085	.047	.126	.176	.170	.138
.300	.125	.183	.246	.230	.294	.348	.336	.304
.450	.313	.358	.408	.407	.456	.515	.500	.464
.600	.500	.533	.569	.577	.612	.679	.660	.618
.750	.688	.708	.731	.741	.763	.838	.817	.766
.900	.875	.883	.892	.899	.907	.993	.971	.908

N = 20, p = 2

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	.000	.050	.103	.074	.128	.168	.165	.137
.300	.176	.218	.261	.248	.293	.332	.327	.299
.450	.353	.385	.419	.418	.452	.494	.487	.458
.600	.529	.553	.578	.583	.608	.654	.644	.611
.750	.706	.721	.736	.743	.759	.810	.799	.761
.900	.882	.888	.894	.899	.905	.963	.952	.906

N = 25, p = 2

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	.034	.073	.113	.090	.131	.163	.162	.137
.300	.205	.236	.270	.259	.293	.325	.321	.298
.450	.375	.400	.426	.425	.451	.484	.479	.455
.600	.545	.564	.583	.587	.605	.641	.635	.608
.750	.716	.727	.739	.745	.756	.795	.789	.758
.900	.886	.891	.896	.899	.904	.948	.941	.904

Table 1. Continued

N = 30, p = 2

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	.056	.087	.120	.100	.133	.161	.160	.138
.300	.222	.248	.275	.266	.293	.320	.318	.297
.450	.389	.409	.430	.429	.450	.477	.474	.453
.600	.556	.570	.586	.589	.604	.633	.629	.606
.750	.722	.731	.741	.746	.755	.786	.782	.757
.900	.889	.893	.896	.899	.903	.939	.934	.904

N = 10, p = 3

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.417	-.275	-.093	-.202	-.031	.012	.010	-.005
.300	-.167	-.050	.100	.040	.178	.250	.222	.198
.450	.083	.175	.293	.270	.374	.477	.428	.390
.600	.333	.400	.486	.487	.560	.692	.627	.571
.750	.583	.625	.679	.690	.734	.897	.819	.741
.900	.833	.850	.871	.880	.898	1.000	1.000	.900

N = 15, p = 3

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.159	-.082	.008	-.049	.039	.087	.083	.052
.300	.045	.109	.183	.154	.225	.278	.267	.235
.450	.250	.300	.358	.349	.403	.464	.448	.412
.600	.455	.491	.533	.537	.575	.645	.624	.581
.750	.659	.682	.708	.717	.740	.821	.797	.744
.900	.864	.873	.883	.889	.898	.992	.966	.900

N = 20, p = 3

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.063	-.009	.050	.012	.070	.109	.107	.078
.300	.125	.169	.218	.198	.246	.286	.280	.253
.450	.313	.347	.385	.380	.416	.459	.451	.422
.600	.500	.525	.553	.556	.582	.630	.620	.586
.750	.688	.703	.721	.727	.743	.797	.785	.745
.900	.875	.881	.888	.893	.899	.961	.948	.900

N = 25, p = 3

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.012	.029	.073	.044	.087	.120	.118	.094
.300	.167	.200	.236	.222	.257	.290	.286	.263
.450	.345	.371	.400	.396	.424	.457	.452	.428
.600	.524	.543	.564	.566	.586	.622	.616	.589
.750	.702	.714	.727	.732	.745	.785	.778	.746
.900	.881	.886	.891	.894	.899	.945	.938	.900

Table 1. Continued

N = 30, p = 3

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	.019	.052	.087	.064	.098	.126	.125	.104
.300	.192	.219	.248	.237	.265	.292	.290	.269
.450	.365	.387	.409	.406	.428	.456	.453	.432
.600	.538	.554	.570	.573	.589	.618	.614	.591
.750	.712	.721	.731	.736	.746	.778	.773	.747
.900	.885	.888	.893	.895	.899	.936	.931	.900

N = 10, p = 4

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.700	-.530	-.275	-.479	-.268	-.285	-.240	-.236
.300	-.400	-.260	-.050	-.176	-.008	.029	.025	.017
.450	-.100	.010	.175	.109	.236	.326	.281	.255
.600	.200	.280	.400	.376	.465	.606	.528	.479
.750	.500	.550	.625	.625	.679	.870	.766	.687
.900	.800	.820	.850	.856	.877	1.000	.995	.880

N = 15, p = 4

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.275	-.190	-.082	-.165	-.067	-.024	-.023	-.053
.300	-.050	.020	.109	.062	.140	.191	.183	.152
.450	.175	.230	.300	.279	.340	.401	.384	.349
.600	.400	.440	.491	.488	.531	.604	.581	.537
.750	.625	.650	.682	.688	.714	.801	.773	.717
.900	.850	.860	.873	.878	.888	.991	.961	.890

N = 20, p = 4

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.133	-.077	-.009	-.060	.003	.042	.041	.012
.300	.067	.113	.169	.141	.192	.232	.227	.199
.450	.267	.303	.347	.336	.375	.419	.411	.381
.600	.467	.493	.525	.525	.553	.603	.592	.557
.750	.667	.683	.703	.708	.725	.782	.769	.728
.900	.867	.873	.881	.885	.892	.958	.944	.893

N = 25, p = 4

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.063	-.020	.029	-.007	.039	.071	.070	.045
.300	.125	.160	.200	.181	.218	.251	.248	.224
.450	.313	.340	.371	.365	.394	.428	.423	.398
.600	.500	.520	.543	.544	.565	.602	.596	.568
.750	.688	.700	.714	.719	.732	.773	.766	.733
.900	.875	.880	.886	.889	.894	.942	.935	.895

Table 1. Continued

N = 30, p = 4

ρ^2	Smith	Ezekiel	Wherry	Olkin-Pratt	Claudy	Pratt	Herzberg	Walker
.150	-.020	.014	.052	.024	.060	.088	.088	.066
.300	.160	.188	.219	.205	.234	.262	.259	.239
.450	.340	.362	.387	.382	.405	.433	.429	.408
.600	.520	.536	.554	.555	.572	.602	.597	.574
.750	.700	.710	.721	.725	.735	.769	.764	.737
.900	.880	.884	.888	.891	.895	.933	.928	.896

Table 2 depicts adjusted R^2 Walker's bias properties or the error that results when estimating ρ^2. Because Walker has similar properties as the Olkin and Pratt formula, the following bias formula presented by Lucke and Embretson (1984) was modified:

$$\text{Bias } R^2_{DW} = 1 - \frac{N - 4.15}{N + 1} * R^2 * \frac{2(1 - R^2)}{(N - 1)} \quad (11)$$

The bias properties for this shrinkage formula show that it is a function of sample size. As would be anticipated, when the sample increases, the bias in this estimator decreases. This formula's bias properties are similar in comparison to other estimators found by Lucke and Embretson (1984).

Table 2. Bias Properties for Adjusted R^2 Walker, N = 10 to 30

ρ^2	N	Bias
.150	10	.174
.300	10	.131
.450	10	.093
.600	10	.061
.750	10	.033
.900	10	.012
.150	15	.109
.300	15	.080
.450	15	.055
.600	15	.034
.750	15	.018
.900	15	.006
.150	20	.079
.300	20	.057
.450	20	.038
.600	20	.023
.750	20	.011
.900	20	.003
.150	25	.062
.300	25	.044
.450	25	.029
.600	25	.017
.750	25	.008
.900	25	.002
.150	30	.051
.300	30	.036
.450	30	.024
.600	30	.014
.750	30	.006
.900	30	.002

Table 3 illustrates Walker's accurateness via standard error estimates for every situation presented in the research. A bootstrapping program conducted 500 resamples to derive the standard error estimate terms presented. Replications of 500 were chosen because the standard error estimates converged quickly at this level and there were relatively no precision differences above this value. As would be expected, bias was greatest under conditions of small N, specifically when N = 10 and 15, where error ranged from 1% to 1.5% in these two situations regardless of p. When N = 20, 25, and 30, standard errors were all < 1%. For instance, Figure 1 shows that the Walker formula produced almost no bias under the extreme case of N = 10, p = 2, and ρ^2 = .15, and became more accurate in this same situation when the sample size increased to N = 15. Further, Figure 2 illustrates this same small bias propensity with the Walker formula, and also the Claudy formula, when p = 2 and ρ^2 = .45, and shows that both the Pratt and Herzberg formulas in this same situation produced overestimations of the ρ^2 value.

Table 3. Standard Error Estimates for Adj. R^2 Walker

p = 2

N	SE	SE Range (Min/Max)
10	.015	(.000, .026)
15	.010	(.000, .017)
20	.008	(.000, .013)
25	.006	(.000, .011)
30	.005	(.000, .009)

p = 3

N	SE	SE Range (Min/Max)
10	.014	(.000, .024)
15	.010	(.000, .017)
20	.007	(.000, .013)
25	.006	(.000, .011)
30	.005	(.000, .009)

p = 4

N	SE	SE Range (Min/Max)
10	.015	(.000, .026)
15	.010	(.000, .016)
20	.007	(.000, .013)
25	.006	(.000, .011)
30	.005	(.000, .009)

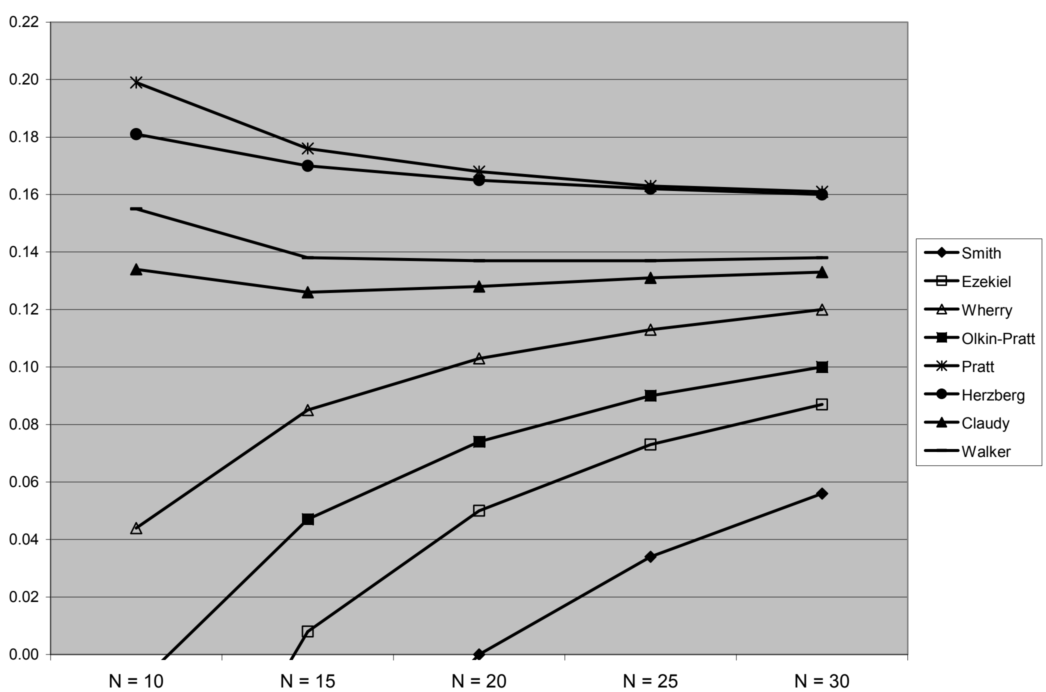

Figure 1. A Comparison of Shrinkage Formulas when $\rho^2 = .15$, p = 2

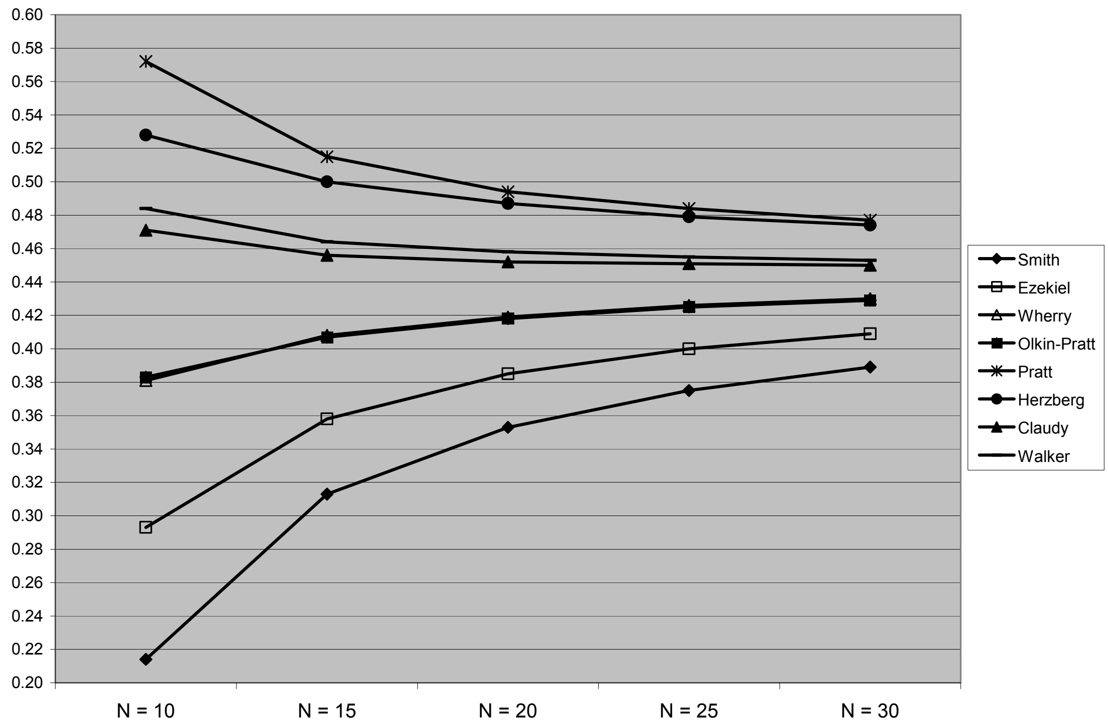

Figure 2. A Comparison of Shrinkage Formulas when $\rho^2 = .45$, p = 2

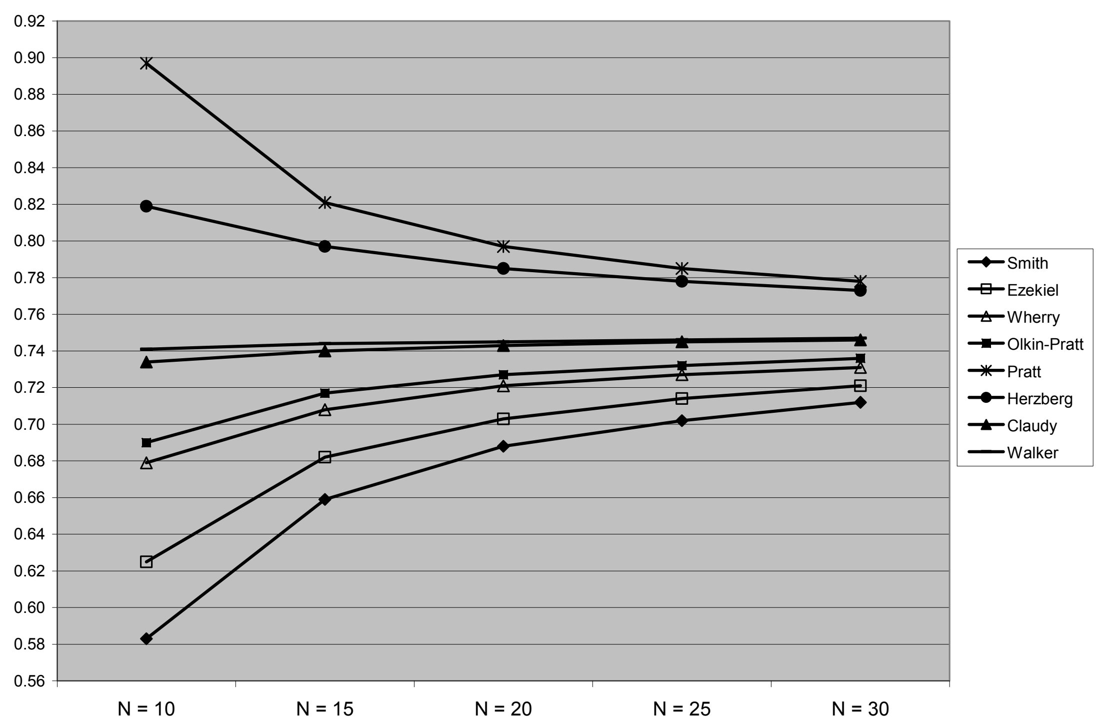

Figure 3. A Comparison of Shrinkage Formulas when $\rho^2 = .75$, p = 3

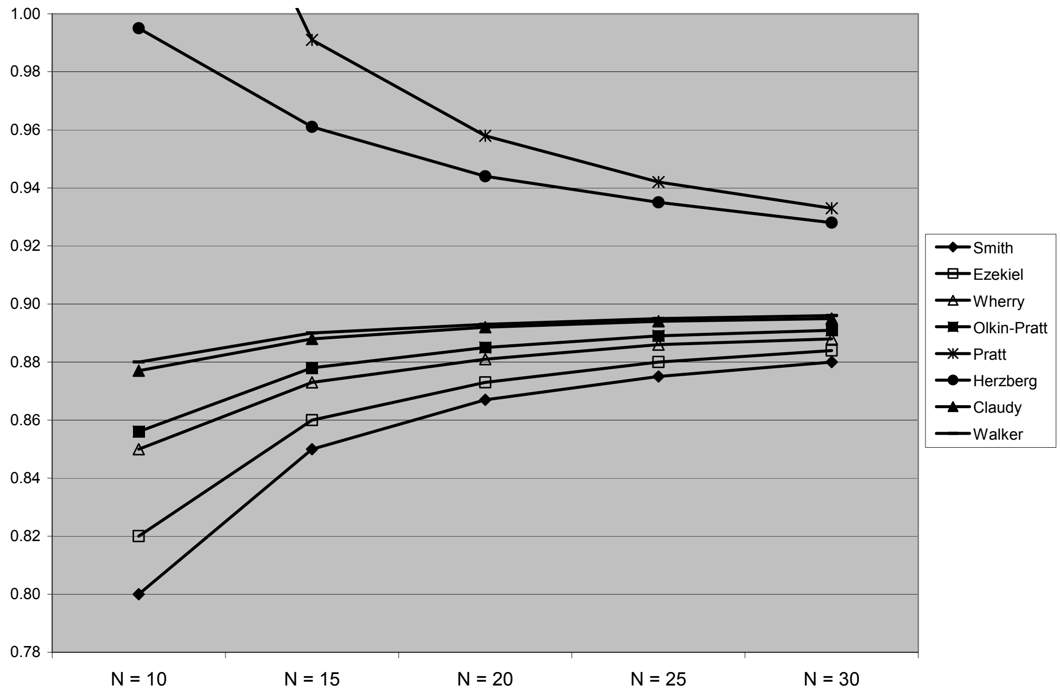

Figure 4. A Comparison of Shrinkage Formulas when ρ^2 = .90, p = 4

Considering data depicted in Figures 3 and 4, it is recommended that when N = 10 to 30 with either p = 3 or 4, use the Walker formula, which was more accurate in every instance than Claudy, and the majority of the time more exact than either Pratt or Herzberg due to their overestimations typically at ρ^2 values of .60, .75., and .90. When N = 10 to 30 and p = 2, the Claudy formula was more accurate than Walker, except in the case where ρ^2 = .15. It is not recommended, however, to use either Smith or Ezekiel in any of the presented situations when $\rho^2 \leq .60$. Wherry and Olkin and Pratt may be regarded in some instances when ρ^2 = .60, but tend to be more accurate in all cases at the .75 and .90 levels.

Lastly, extreme research situations can produce adjusted R^2 values that are nonsensical. For example, the negative values depicted in Table 1 and Figure 1 have been noted before in previous research associated with shrinkage formulas by Huberty and Mourad (1980), where it was found that, “Negative values will result from using a small R^2 value and/or a small N/p ratio” (p. 108). Thus, these negative figures should be considered to take on the value of zero.

Conclusion

When estimating the population multiple correlation coefficient, reducing the positive bias found in R^2, the coefficient of determination, is approached via an unbiased estimator called the adjusted R^2. However, a caveat with adjusted R^2 is that not all unbiased estimators of ρ^2 function the same under varying research situations. The goal of this research was to look at this issue and determine which of the eight estimators chosen performed the most consistently under biased research conditions often found within the field of educational research, where N was small and the number of X variables ranged from 2 to 4.

The results of this study yielded no definitive answers pertaining to the best estimators in every situation examined, but it did ascertain that the two most consistently accurate formulas in the many conditions studied were Claudy and Walker. The tabled data derived from this research should provide researchers and students with information to understand when to use various adjusted R^2 estimators pertaining to a given research situation. Also, this research introduced a new shrinkage formula, Adj. R^2_{DW}, and provided a complete error profile and comparison analysis under extreme research conditions for the user's consideration. Future research affiliated with shrinkage formulas should include the performance of these eight estimators under the same extreme conditions, but when operating in very biased distributional situations such as with outlier data points and/or under non-normal conditions of various skew.

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed). Upper Saddle River, NJ: Prentice Hall.

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journalof Mathematical and Statistical Psychology, 28*, 79-87.

Carter, D. S. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. *Educational and Psychological Measurement, 39*, 261-266.

Claudy, J. G. (1972). A comparison of five variable weighting procedures. *Educational and Psychological Measurement, 32*, 311-322.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *AppliedPsychological Measurement, 2*, 595-607.

Ezekiel, M. (1929). The application of the theory of error to multiple and curvilinear correlation. *American Statistical Association Journal, 24*, 99-104.

Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.

Herzberg, P. A. (1969). The parameters of cross-validation. *Psychometric Monograph,16*.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement, 40*, 101-112.

Lucke, J. F., & Embretson, S. (1984). The biases and mean squared errors of estimators of multinormal squared multiple correlation. *Journal of Educational Statistics, 9*, 183-192.

Olkin, I., & Pratt, J. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29*, 201-211.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Schumacker, R. E., Mount, R. E., & Monahan, M. P. (2002). Factors affecting multiple regression and discriminant analysis with a dichotomous dependent variable: Prediction, explanation, and classification. *Multiple Linear Regression Viewpoints, 28*, 32-39.

Walker, D. A. (2006, April). *A comparison of eight shrinkage formulas under extreme conditions.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2*, 440-451.