Correcting for Range Restriction When the Population Variance is Unknown

Ralph A. Alexander, George M. Alliger, and Paul J. Hanges
The University of Akron

Correction of correlations diminished by range restriction is a commonly suggested psychometric technique. Such corrections may be appropriate in applied settings, such as educational or personnel selection, or in more theoretical applications, such as meta-analysis. However, an important limitation on the practice of range restriction corrections exists—an estimate of the unrestricted population variance is required. This article outlines and examines the accuracy of a method for estimating the unrestricted variance of a variable from the restricted sample itself. This method is based on the observation that it is possible to table a function of the truncated normal distribution that will allow the extent or point of truncation to be estimated (Cohen, 1959). The correlation of the truncated variable with other variables may then be corrected by standard restriction of range formulas. The method also allows for correction of the mean of the restricted variable.

Since the introduction by Thorndike (1947) of Pearson's selection formulas into psychometric theory and practice, the phenomenon of correlations reduced by range restriction has received a considerable amount of attention (e.g., Alexander, Barrett, Alliger, & Carson, 1984; Alexander, Carson, Alliger, & Barrett, 1984; Brewer & Hills, 1969; Greener & Osburn, 1979; Gross, 1982; Gross & Fleischman, 1983; Linn, 1968). It has been determined that corrected correlations are less biased than the uncorrected estimates over a wide range of assumption violations (Linn, 1983), and application of correction formulas is now recognized professional practice (American Psychological Association, 1980).

Of the three cases of range restriction examined by Thorndike (1947), Case 2 is addressed here, which assumes knowledge of the unrestricted variance in the restricted, or truncated, variable. Both Case 1 and Case 2 deal with direct restriction on the predictor or criterion, but Thorndike (1949) regarded Case 2 as far more common of the two in actual practice. Case 3 addresses indirect restriction of range and is beyond the scope of this paper.

The Case 2 formula requires that the unrestricted variance of a variable be known in order to correct for restriction on that variable. If the unrestricted predictor variance is not observed, some estimate of it must be obtained. Often, for example, a test may be used for selection but no data are available on the variance in the unrestricted population. Also, it is commonly the case that summary data sets include only information on the selected sample(s). As a result, the researcher who wishes to correct for range restriction, whether for a single study or as part of a meta-analysis, must estimate the unrestricted variance

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 4, Fall 1984, pp. 431–437

0146-6216/84/040431-07$1.60

on the selector test. Usually such estimates are obtained from a review of literature relevant to the predictor test (e.g., from published normative data on the test).

This article presents a method for obtaining unrestricted estimates of univariate variances that have been reduced by direct truncation, assuming only a normal distribution in the untruncated sample. These estimates may then be used in the standard Case 2 range restriction correction formula. The method consists of using a simple tabled function that estimates the degree of the truncation occurring in a sample. Once the degree of truncation has been estimated, other characteristics of the sample such as the variance and mean can be estimated.

Method

Cohen (1959) proposed the ratio $SD^2/(\bar{X} - X_c)^2$, the sample variance over the squared difference between the sample mean and point of truncation. This ratio is useful because a normal distribution, truncated at X_c, will have a unique value of $SD^2/(\bar{X} - X_c)^2$. Although other functions of the standard score formula would also give a unique value for each degree of truncation, Cohen's ratio has the advantage of ease of calculation from the sample data. Two normal functions, the standard deviation in standardized form in a truncated distribution and the z-score representing the truncation point, are tabled directly against Cohen's ratio (see Table 1). These tabled values can then easily be used to correct means, standard deviations, and correlations.

The sample of interest is assumed to be a random sample from a normal population truncated at some unknown point. Cohen's ratio is calculated from the sample mean and standard deviation ($\bar{X}$, SD) and the lowest observed variate-value (X_c). (In this discussion, truncation is assumed to occur at the lower end of the distribution. If truncation has occurred at the upper end of a distribution, the method described is the same except that X_c is taken as the highest observed sample value and the algebraic sign of the z-score of Table 1 is reversed.) Once Cohen's ratio has been calculated, the closest tabled value (Table 1) is located. The corresponding z-cut, that is, the z-score identifying the point of truncation, and the standard deviation in a normal distribution truncated at that point may then be obtained from Table 1. Since the standard deviation in Table 1 is the standardized value of the standard deviation after truncation (with a nontruncated value of 1.0), that tabled value also represents the proportional reduction in the standard deviation due to range restriction. The observed sample standard deviation may be corrected by:

$$\hat{SD} = SD_{obs}/SD_{tab} \quad , \tag{1}$$

where SD_{obs} is the observed standard deviation in a restricted sample, and SD_{tab} is the tabled restricted normal standard deviation. In effect, Equation 1 increases SD_{obs} proportionately to the amount of truncation estimated by Cohen's ratio.

The corrected mean may also then be estimated by:

$$\hat{\bar{X}} = X_c - z\hat{SD} \quad , \tag{2}$$

where X_c is the observed truncation point, that is, the lowest observed value of the truncated variable. The value z in Equation 2 is the truncation point in standard score form as obtained from Table 1, and $\hat{SD}$ is the corrected standard deviation obtained from Equation 1. These formulas are identical in result to Cohen's formulas but more easily used.

For correcting correlations when the predictor has been truncated, it is, of course, $\hat{SD}$ obtained from Equation 1 that is pertinent. Using the corrected standard deviation, Thorndike's Case 2 formula becomes:

$$\hat{r} = r/\{r^2 + [\hat{SD}^2/SD^2_{obs}] - [\hat{SD}^2/SD^2_{obs}](r^2)\}^{1/2} \quad , \tag{3}$$

Table 1
Cohen's Ratio and Corresponding Retricted Standard Deviation of z Score for Truncation Point

$SD^2/(\bar{X}-x_c)^2$	SD_{tab}	z	$SD^2/(\bar{X}-x_c)^2$	SD_{tab}	z	$SD^2/(\bar{X}-x_c)^2$	SD_{tab}	z
.109	.993	-3.00	.389	.784	-0.95	.734	.433	1.10
.113	.992	-2.95	.399	.775	-0.90	.740	.427	1.15
.116	.991	-2.90	.409	.765	-0.85	.746	.421	1.20
.120	.990	-2.85	.419	.756	-0.80	.751	.415	1.25
.124	.989	-2.80	.429	.746	-0.75	.757	.409	1.30
.128	.987	-2.75	.438	.736	-0.70	.762	.403	1.35
.132	.986	-2.70	.448	.726	-0.65	.767	.398	1.40
.137	.984	-2.65	.458	.717	-0.60	.772	.392	1.45
.141	.982	-2.60	.468	.707	-0.55	.777	.387	1.50
.146	.980	-2.55	.477	.697	-0.50	.782	.381	1.55
.151	.978	-2.50	.487	.688	-0.45	.787	.376	1.60
.156	.975	-2.45	.497	.678	-0.40	.791	.371	1.65
.161	.972	-2.40	.506	.668	-0.35	.796	.366	1.70
.167	.969	-2.35	.516	.659	-0.30	.800	.361	1.75
.172	.966	-2.30	.525	.649	-0.25	.804	.356	1.80
.178	.963	-2.25	.534	.640	-0.20	.809	.352	1.85
.184	.959	-2.20	.544	.630	-0.15	.813	.347	1.90
.190	.955	-2.15	.553	.621	-0.10	.817	.343	1.95
.197	.951	-2.10	.562	.612	-0.05	.820	.338	2.00
.203	.946	-2.05	.571	.603	0.00	.825	.334	2.05
.210	.942	-2.00	.580	.594	0.05	.828	.329	2.10
.217	.936	-1.95	.588	.585	0.10	.832	.325	2.15
.224	.931	-1.90	.597	.576	0.15	.835	.321	2.20
.231	.926	-1.85	.605	.568	0.20	.838	.317	2.25
.239	.920	-1.80	.614	.559	0.25	.842	.313	2.30
.247	.914	-1.75	.622	.551	0.30	.845	.309	2.35
.254	.907	-1.70	.630	.542	0.35	.848	.306	2.40
.263	.901	-1.65	.638	.534	0.40	.851	.302	2.45
.271	.894	-1.60	.646	.526	0.45	.855	.298	2.50
.279	.886	-1.55	.653	.518	0.50	.857	.295	2.55
.288	.879	-1.50	.661	.510	0.55	.860	.291	2.60
.296	.871	-1.45	.668	.503	0.60	.864	.288	2.65
.305	.863	-1.40	.675	.495	0.65	.867	.285	2.70
.314	.855	-1.35	.682	.488	0.70	.869	.281	2.75
.323	.847	-1.30	.689	.481	0.75	.868	.278	2.80
.332	.838	-1.25	.696	.473	0.80	.875	.275	2.85
.342	.830	-1.20	.703	.466	0.85	.875	.272	2.90
.351	.821	-1.15	.709	.460	0.90	.882	.269	2.95
.361	.812	-1.10	.716	.453	0.95	.882	.266	3.00
.370	.803	-1.05	.722	.446	1.00			
.380	.794	-1.00	.728	.440	1.05			

where r is the observed correlation,

SD_{obs} is the observed standard deviation, and

$\hat{SD}$ is the corrected standard deviation from Equation 1.

In order to test the performance of Equation 3 when using estimates of corrected standard deviations according to the method suggested above, both computer simulations and the examination of real data

sets were employed. Using the International Mathematical and Statistical Library (IMSL, 1979), simulations were carried out in the following manner. Bivariate normal data sets consisting of 5,000 samples of $n = 30$ and 60 and 2,000 samples of $n = 100$ were generated for $\rho = .2, .4, .6$, and .8 and truncation points of $z = -2, -1.5, -1, -.5, 0, +.5$, and $+1$. For each simulated sample, the observed reduced mean, standard deviation, and lowest observed value in the truncated variable and the reduced correlation between variables were found. The corrected values of these parameters were then calculated using the above formulas, and the accuracy of the results was noted.

A similar examination was carried out on two large sets of real data. A sample of 1,235 cases consisting of ACT scores and first-year college grade point averages (GPA) and a sample of 675 cases of ACT scores and scores on a computer aptitude test were used. Each sample was truncated at 0 (i.e. at the mean), -1 and -2 standard deviations, and the observed and corrected standard deviations, means, and correlations were found.

Results

Results from the simulation are summarized in Table 2. Shown in Table 2 are the mean observed and corrected correlations for different ρ and points of truncation. Since the results were not affected by n, Table 2 is based on simulations of $n = 60$. Note that even though the unrestricted variance is an estimate using Table 1 and Equations 1 and 3, in every case the corrected correlation is closer to the original value than is the uncorrected value except for the most minimal cut (-2.0 SD). The corrected correlations are overestimates except for fairly severe truncation ($+1.0$ SD), but the overestimation is never greater than .02.

In addition to the accuracy of the method in recapturing the unrestricted correlation, the variance in the corrected values is also of interest. Table 3 presents (from the simulations) the standard deviation of the restricted uncorrected correlations, the sample coefficients corrected by the usual Thorndike method when the unrestricted population variance is assumed known, and those corrected by the method presented in this article when the variance in the restricted variable is estimated from the sample data.

Inspection of Table 3 shows that the standard deviations of the corrected rs are generally larger than those for restricted but uncorrected coefficients. Further, corrected coefficients using the method of this article have larger variance than those corrected under the usual method. This is to be expected, since under the present method an additional parameter (the unrestricted variance) is being estimated from

Table 2
Computer Simulation Results: Comparison of Mean Observed and Corrected Correlations for Differing ρ and Truncation Point

Truncation Point (SD)	True ρ .20	.40	.60	.80
-2.0	.19 (.21)	.38 (.42)	.58 (.62)	.78 (.81)
-1.5	.18 (.21)	.36 (.41)	.55 (.62)	.76 (.81)
-1.0	.16 (.21)	.33 (.41)	.51 (.61)	.72 (.81)
-0.5	.14 (.21)	.29 (.41)	.46 (.61)	.68 (.81)
μ	.12 (.21)	.25 (.42)	.41 (.61)	.62 (.81)
+0.5	.10 (.21)	.22 (.41)	.36 (.61)	.56 (.80)
+1.0	.09 (.19)	.19 (.39)	.31 (.59)	.50 (.79)

Note. Corrected values are in parentheses. Data based on an n-size of 60 with 5,000 iterations.

Table 3
Standard Deviations of Restricted Uncorrected Correlations and Correlations Corrected Using Known Variance and Estimated Variance

Cuts on X* (σ)	ρ .2			.4			.6			.8		
	n=30	n=60	n=100	n=30	n=60	n=100	n=30	n=60	n=100	n=30	n=60	n=100
-2.0												
(1)	177	124	097	158	110	086	124	086	067	074	050	039
(2)	189	131	102	162	111	086	116	079	061	058	040	031
(3)	206	137	105	177	119	092	132	091	070	073	051	040
-1.5												
(1)	178	124	098	160	112	088	129	090	070	079	054	042
(2)	202	139	109	172	118	092	123	084	066	060	041	032
(3)	220	147	113	189	128	099	142	098	076	079	056	043
-1.0												
(1)	180	126	100	165	115	091	138	100	075	091	063	047
(2)	222	155	122	212	131	102	137	092	071	066	044	034
(3)	245	164	128	190	143	111	161	110	085	091	063	048
-0.5												
(1)	182	128	100	171	120	093	148	103	079	105	072	054
(2)	250	176	138	216	149	115	157	105	080	074	050	037
(3)	283	189	145	246	167	126	189	129	097	109	075	056
μ												
(1)	182	129	101	174	124	096	157	112	086	120	084	064
(2)	283	203	159	247	173	135	183	124	094	089	058	044
(3)	323	224	172	282	199	152	217	158	121	129	094	072
+0.5												
(1)	188	129	101	182	125	098	170	116	090	139	095	073
(2)	328	232	182	293	200	155	226	145	109	113	067	050
(3)	367	260	202	327	231	182	258	183	146	157	112	089
+1.0												
(1)	185	131	102	181	128	099	173	122	094	149	104	080
(2)	362	266	209	328	233	179	261	172	127	137	080	058
(3)	389	292	231	351	260	207	284	207	166	181	129	102

Note. Based on 5,000 replications for n=30 and n=60 and 2,000 replications for n=100. Decimals are omitted.
*(1) = standard deviation of range restricted, uncorrected r.
(2) = standard deviation of corrected r assuming unrestricted population variance (σ_X) is known.
(3) = standard deviation of corrected r assuming unrestricted population variance (σ_X) is not known.

sample data. The results in Table 3 also show, however, that the difference in variance between the two correction methods is sufficiently small that restricted correlations corrected by this method will not have unacceptable wide confidence intervals relative to the usual correction method.

Table 4 contains the results for the correction formulas on the two sample data sets. For Sample 1, the unrestricted correlation between ACT and GPA is .53. Correcting the restricted correlations results in an overcorrection under slight truncation (-2 SD) and an undercorrection under more stringent truncation. For Sample 2, the unrestricted correlation between ACT and computer aptitude scores is .62. Corrected correlations overestimate under each degree of truncation. Overcorrections in both samples range from .03 to .16 (Sample 2, truncation at the mean). Although these corrections are not perfectly accurate, it should be noted that, for both studies, they diverge from corrections using actual knowledge of the unrestricted variance by only $\pm .04$.

Discussion

The results of both the simulations and the examination of the real data sets indicate that correction for correlations when the actual unrestricted variance is unknown may be useful. The accuracy of correction does not appear to be different than that which others have noted on actual truncated data, using known unrestricted variance (Lee, Miller, & Graham, 1982). Reporting of such corrected correlations should be done cautiously, and reporting the uncorrected correlations along with the corrected values is good practice whenever such corrections are used. In any case, it can be said that the corrections are likely to be more accurate than the uncorrected values.

Several avenues of further research would be useful. These include an examination of the way in which violations of normality and linearity affect correction of correlations using an estimate of the unrestricted variance as outlined in this paper. Application of this correction method to additional real data would also appear useful in order to further gauge the degree to which confidence can be placed in this method. Beyond this, it should be noted that very often restriction of range is something other than strict truncation; in personnel selection, for example, some applicants above the cutoff score may be rejected and some below accepted. The performance of this correction method under such incomplete truncation is also unknown and might profitably be examined.

Table 4
Comparison of Observed and Corrected Values of $\bar{X}$, SD, and r for Varying Degrees of Truncation for Two Student Samples

Truncation	N	$\bar{X}$	SD	r
None				
Sample 1	1235	18.14	5.24	.53
Sample 2	675	18.72	5.21	.62
-2σ				
Sample 1	1220	18.28 (17.77)	5.12 (5.57)	.52 (.55)
Sample 2	663	18.94 (18.39)	5.01 (5.52)	.61 (.65)
-1σ				
Sample 1	1006	19.84 (18.46)	4.17 (5.20)	.41 (.49)
Sample 2	562	20.76 (17.53)	4.22 (5.88)	.54 (.66)
μ				
Sample 1	616	22.53 (20.55)	2.91 (4.12)	.30 (.41)
Sample 2	353	22.81 (15.29)	3.11 (6.18)	.53 (.78)

Note. Corrected values are in parentheses.

References

Alexander, R. A., Barrett, G. V., Alliger, G. M., & Carson, K. P. (1984). *Toward a general model of nonrandom sampling and the impact on population correlation: Generalizations of Berkson's fallacy and restriction of range*. Manuscript submitted for publication.

Alexander, R. A., Carson, K. P., Alliger, G. M., & Barrett, G. V. (1984). Correction for restriction of range when both *X* and *Y* are truncated. *Applied Psychological Measurement, 8,* 231–241.

American Psychological Association, Division of Industrial-Organizational Psychology. (1980). *Principles for the validation and use of personnel selection procedures: Second edition*. Dayton OH: Author.

Brewer, J. K., & Hills, J. R. (1969). Univariate selection: The effects of size of correlation, degree of skew, and degree of restriction. *Psychometrika, 26,* 347–372.

Cohen, A. C. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics, 1,* 217–237.

Greener, J. M., & Osburn, H. G. (1979). An empirical study of the accuracy of corrections for restriction in range due to explicit selection. *Applied Psychological Measurement, 3,* 31–41.

Gross, A. L. (1982). Relaxing assumptions underlying corrections for restriction of range. *Educational and Psychological Measurement, 42,* 795–801.

Gross, A. L., & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. *Applied Psychological Measurement, 7,* 227–237.

IMSL Library. (1979). Houston TX: International Mathematical and Statistical Libraries.

Lee, R., Miller, K. J., & Graham, W. Y. (1982). Corrections for restriction of range and attenuation in criterion-related studies. *Journal of Applied Psychology, 67,* 637–639.

Linn, R. L. (1968). Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin, 69,* 69–73.

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20,* 1–15.

Thorndike, R. L. (1947). *Research problems and techniques* (Report No. 3). Washington DC: AAF Aviation Psychology Program Research Reports, U.S. Government Printing Office.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

Author's Address

Send requests for reprints or further information to Ralph A. Alexander, Department of Psychology, The University of Akron, Akron OH 44325, U.S.A.