An equation was derived to determine the relationship between the pooled within-subgroup $r$ (correlation coefficient) and the $r$ obtained from the total group data. It was, thus, possible to assess the amount of distortion introduced by pooling heterogeneous subgroups. As a basis for deciding whether to pool two subgroups in order to calculate a single $r$ for the total group, a two-stage procedure was recommended: (1) comparison of the two within-subgroups $r$'s; and (2) comparison of the total group $r$ and the pooled within-subgroup $r$. On the basis of results for the second stage test, distortion in the total group $r$ was shown to be a function of the pattern of subgroup mean differences, total group sample size, and the magnitude of the pooled within-subgroup $r$. Implications were discussed. (Author)
THE EFFECT OF POOLING TWO HETEROGENEOUS SUBGROUPS ON
THE PRODUCT-MOMENT CORRELATION COEFFICIENT

Alan L. Socklof
Temple University

An equation was derived to determine the relationship between the pooled within-subgroup r and the r obtained from the total group data. It was, thus, possible to assess the amount of distortion introduced by pooling heterogeneous subgroups. As a basis for deciding whether to pool two subgroups in order to calculate a single r for the total group, a two-stage procedure was recommended: (1) comparison of the two within-subgroup r's; and (2) comparison of the total group r and the pooled within-subgroup r. On the basis of results for the second stage test, distortion in the total group r was shown to be a function of the pattern of subgroup mean differences, total group sample size, and the magnitude of the pooled within-subgroup r. Implications were discussed.

Frequently, in the psychological and educational literature, correlational studies are reported in which product-moment correlations are calculated between two variables for sets of data pooled across two, possibly heterogeneous, categorical subgroups. The pooling of heterogeneous subgroups to calculate a product-moment correlation was first discussed by Karl Pearson (Pearson, Lee, & Bramley-Moore, 1899). By way of illustration, these authors presented correlational data between length and breadth of skulls for 806 males ($r = .0869$) and 340 females ($r = -.0424$). When the two subgroups were pooled, an $r$ of .1968 was obtained, and this $r$ was considered to represent a large spurious correlation.

From a sampling of recent introductory statistics textbooks in psychology and education, it was found that writers do discuss the effects on the correlation coefficient resulting from pooling heterogeneous subgroups (Games & Klare, 1967; Glass & Stanley, 1970; Guilford, 1965; Walker & Lev, 1969; among others). Where references are made, Dunlap's (1937) paper on the combinative properties of correlation coefficients is most frequently cited. Dunlap presented a method for calculating a total group correlation coefficient from subgroup correlation coefficients and the means and standard deviations of the variables.

To date, a mathematical formulation of the effects on the correlation coefficient from pooling heterogeneous subgroups has been lacking. In lieu of such a formulation, textbook writers have tended to stress cautious interpretation and the use of subgroup correlation coefficients to help provide a rational explanation for correlation results in the total group. The major interest of this paper is the derivation of a mathematical formulation and the demonstration of the effects of
pooling two subgroups on the total group correlation coefficient. Of additional interest is a procedure to guide the decision concerning the pooling of data for the purpose of calculating a single correlation coefficient.

Formulation

Given two subgroups, let \( n_1 \) and \( n_2 \) be the subgroup sample sizes, where \( n_1 + n_2 = N \). Let \( U \) and \( V \) be the distances between the subgroup means for \( X \) and \( Y \), respectively, i.e., \( U = \bar{X}_2 - \bar{X}_1 \) and \( V = \bar{Y}_2 - \bar{Y}_1 \).

Sum of Squares and Cross Products

The sum of cross products for the total group, \( SS(XY)_t \), can be defined

\[
SS(XY)_t = SS(XY)_1 + SS(XY)_2 + UV\left(\frac{n_1 n_2}{N}\right),
\]

where \( SS(XY)_1 \) and \( SS(XY)_2 \) are the within-subgroup sums of cross products. Since the sum of squares for \( X \) in the total group is actually the sum of cross products with respect to itself,

\[
SS(X)_t = SS(X)_1 + SS(X)_2 + U^2\left(\frac{n_1 n_2}{N}\right).
\]

Similarly, for \( Y \),

\[
SS(Y)_t = SS(Y)_1 + SS(Y)_2 + V^2\left(\frac{n_1 n_2}{N}\right).
\]

Total Group \( r_t \)

Using Equations 1, 2, and 3, the correlation coefficient for the total group is:

\[
r_t = \frac{SS(XY)_1 + SS(XY)_2 + UV(n_1 n_2/N)}{\sqrt{(SS(X)_1 + SS(X)_2 + U^2(n_1 n_2/N))(SS(Y)_1 + SS(Y)_2 + V^2(n_1 n_2/N))}}.
\]

An equivalent form of this equation was derived by Dunlap (1937). If \( SS(XY)_w = SS(XY)_1 + SS(XY)_2 \), \( SS(X)_w = SS(X)_1 + SS(X)_2 \), and \( SS(Y)_w = SS(Y)_1 + SS(Y)_2 \), then the equation defining \( r_t \) may be simplified:
The correlation coefficient for the total group is therefore expressed in terms of pooled within-subgroup sums of squares and cross products. subgroup sample sizes, and distances between the subgroup means.

Pooled Within-Subgroup \( r_w \)

The pooled within-subgroup correlation coefficient is obtained from pooling of the subgroup sums of squares and cross products:

\[
\begin{align*}
\hat{r}_w &= \frac{SS(XY)_{11} + SS(XY)_{21}}{\sqrt{(SS(X)_{11} + SS(X)_{21})(SS(Y)_{11} + SS(Y)_{21})}} \\
&= \frac{SS(XY)_{11}}{\sqrt{SS(X)_{11} + SS(Y)_{11}}}. 
\end{align*}
\]

It should be clear from inspection of Equation 5 that for equal variances of \( X \) in both subgroups and equal variances of \( Y \) in both subgroups, \( r_w \) is a weighted arithmetic mean of the within-subgroup correlations, weighted by the number of observations in each subgroup.

Furthermore, \( r_w \) may be compared to \( r_t \) in two ways. First, \( r_w \) is a special case of \( r_t \) resulting when subgroup mean differences are non-existent (i.e., \( U = V = 0 \)). Second, \( r_w \) is that special case of \( r_t \) when subgroup differences are eliminated statistically. The latter comparison requires the form of a first-order partial correlation \( r_{X_tY_tZ} \), where \( X_t \) and \( Y_t \) are the two variables measured in the total group and \( Z \) is a dichotomous variable indicating subgroup membership. In the formula for the first-order partial correlation, since \( r_{X_tZ} \) and \( r_{Y_tZ} \) are point-biserial correlations, the result of operating upon the formula for the first-order partial correlation is:

\[
\begin{align*}
\hat{r}_{X_tY_tZ} &= \frac{SS(XY)_{11} - UV(n_1n_2/N)}{\sqrt{(SS(X)_{11} - U^2(n_1n_2/N))(SS(Y)_{11} - V^2(n_1n_2/N))}}. 
\end{align*}
\]

Since the above equation represents an alternative definition of \( r_w \), then \( r_w \), the pooled within-subgroup correlation coefficient, is also
the result of statistically eliminating subgroup differences from the
total group correlation coefficient. The latter definition of \( r_w \) sug-
gests that \( r_w \) can be meaningfully used as a descriptive statistic with
a known sampling distribution.

**Further Derivation of \( r_t \)**

If the numerator and denominator of Equation 4 are each divided by
the product of the pooled within-subgroup standard errors of the mean
(\( s_{xw} \) and \( s_{yw} \)), a more convenient definition of \( r_t \) arises. To complete
this series of operations, by defining \( t_x = U/s_{xw} \) and \( t_y = V/s_{yw} \) as
subgroup mean differences measured in units of standard errors, the
following final form results:

\[
\frac{(N - 2) r_w + t_x t_y}{\sqrt{(N - 2) + t_x^2}(N - 2) + t_y^2}}
\]

In this form, \( r_t \) is defined in terms of total group sample size, the
pooled within-subgroup correlation coefficient, and subgroup mean
differences that are measured in units distributed as Student's t under
the appropriate assumptions.

**The Decision to Pool**

The following is a simple two-stage procedure, recommended as a
basis for the decision to pool two subgroups in order to calculate a
single correlation coefficient for the total group data.

1. The two within-subgroup correlation coefficients should be
   compared. Under the hypothesis \( H_0: \rho_1 = \rho_2 \), the unit-normal \( z \) test,
   relying on Fisher's \( r-z \) transformation (\( z_F \)), can be used:

\[
z = \frac{z_{F1} - z_{F2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}
\]

If \( H_0 \) is rejected, it is unreasonable to calculate \( r_w \) as a measure of
correlation for the two subgroups. If $H_0$ is accepted, then $r_w$ can be considered a useful measure of correlation for the two subgroups, and the second stage should be followed.

2. In order to assess the distortion introduced by pooling the two subgroups, $r_t$ should be compared to $r_w$ under the hypothesis $H_0$: $p_t = p_w$. For this test, if $r_w$ is considered an estimate of $p_w$,

$$z = \frac{z_{r_t} - z_{r_w}}{\sqrt{\frac{1}{n} - \frac{1}{4}}}$$

If $H_0$ is rejected, then $r_t$ may be considered distorted, but $r_w$ can be used as a measure of correlation for the two subgroups. If $H_0$ is accepted, then pooling the data from the two subgroups in order to calculate a total group correlation coefficient appears to be a parsimonious and reasonable procedure.

**Examples of Distortion**

According to Equation 6, distortion in $r_t$ is affected by subgroup centroid differences and total group sample size. Three patterns of subgroup centroid differences are of interest and are shown in Figure 1.

(1) If the mean of Subgroup 2 is higher than the mean of Subgroup 1 on both variables (Cases a, b, and c of Figure 1), the greater the difference between the subgroups on the two variables, the more exaggerated the value of $r_t$ in a positive direction. (2) If the mean of Subgroup 2 is higher than the mean of Subgroup 1 on one variable, and equal on the other variable (Cases d, e, and f of Figure 1), the greater the difference between the two subgroups on the one variable, the closer $r_t$ is to a value of 0.00. (3) If the mean of Subgroup 2 is higher than
the mean of Subgroup 1 on one variable, and lower on the other variable (Cases g, h, and i of Figure 1), the greater the difference between the subgroups on the two variables, the more exaggerated the value of $r_t$ in a negative direction. Furthermore, for constant differences between the subgroup centroids as measured in standard errors, increasing the total group sample size serves to minimize the effects of subgroup centroid differences, i.e., $r_t$ approaches $r_w$.

The effects of subgroup sample size discrepancy on the calculation of $r_t$ and $r_w$ can be shown by reference to Equations 4 and 5. According to Equation 5, for constant total group sample size, the larger the discrepancy between the subgroup sample sizes, the greater the influence of the larger subgroup in the calculation of $r_w$. In addition, according to Equation 4, for constant differences between subgroup centroids and for constant total group sample size, the larger the discrepancy between $n_1$ and $n_2$, the smaller the effect of subgroup centroid differences in the calculation of $r_t$, and, thus, the more equal the values of $r_t$ and $r_w$.

For the second stage test, in order to demonstrate the amount of distortion introduced by pooling heterogeneous subgroups, Equation 6 was utilized to calculate $r_t$ under varying sample conditions. The sample conditions were derived from combinations of four magnitudes of subgroup centroid differences for the three patterns, five total group sample sizes ($N$), and three values of $r_w$. By assuming two bivariate normal populations, the magnitude of difference between subgroup means can be represented by employing four-decimal critical values of Student's $t$ distribution for $p < .05$, $p < .01$, $p < .001$, and $p < .0001$, obtained for $N-2$ df from Sockloff & Edney's (1972) tables. The .001 and .0001 significance levels were used because these levels represent extreme differences that are sometimes found in research data, although not necessarily
The five total group sample sizes were 10, 50, 100, 200, and 1000. The three values of $r_w$ were .8000, .4000, and 0.0000, chosen to represent high, moderate, and low correlations, respectively.

Tables 1, 2, and 3 present calculated values of $r_t$ derived from the values of $r_w$ under the varying sample conditions. Also included in these tables are the second stage two-tailed tests of $H_0$: $\rho_t = \rho_w$ to assess the amount of distortion introduced under the conditions. Negative values of $r_w$ are not shown in the tables since the effects of the three patterns for negative $r_w$'s are opposite in sign from those shown for positive $r_w$'s, e.g., the amount of exaggeration in a positive direction for a positive correlation under Pattern 1 is equal to the amount of exaggeration in a negative direction for a negative correlation under Pattern 3.

As shown in Table 1, under Patterns 1 and 2, large differences between the subgroup means have a small effect on the value of $r_t$ when $r_w = .8000$. Under both patterns, a total group sample size of 50 appears to be sufficient to minimize the distortion introduced by subgroup mean differences that are significant at the .0001 level. On the other hand, the results were quite different under Pattern 3. When the means of the two subgroups are significantly different at the .0001 level, but in opposite directions, for a total group sample size of 50 $r_t$ was calculated to be .3089, which is significantly different from an $r_w$ of .8000. Furthermore, even for a total group sample size of 1000, the pooling of subgroups when subgroup means differ in opposite directions at the .0001 level produced an $r_t$ of .7729. Although this value of $r_t$ is significantly different from an $r_w$ at the .05 level,
one can argue that such statistically significant differences between
$r_w$ and $r_v$ have little practical significance.

According to Table 2, when $r_w = .4000$, the results for Pattern 1

suggestion that total group sample sizes of 50 are sufficient to avoid
distortions introduced by pooling subgroups when both sets of subgroup
means differ significantly in the same direction at the .0001 level.
Under Pattern 2, significant distortion was not found, even for a total
group sample size of 10. The Pattern 3 results suggest that a total
group sample size of 200 will avoid distortion in the calculation of $r_w$.

According to Table 3, total group sample sizes of 50 appear to be

sufficient to avoid distortion when $r_w = 0.0000$ and the two sets of
subgroup means differ at the .0001 level. Based on the symmetry of the
sampling distributions of $r$ when $p = 0$, this conclusion holds for subgroup
means differing in the same or opposite directions.

Discussion

The various results clearly suggest the varieties of distortion
that may be introduced by haphazardly pooling subgroups of data for the
purpose of calculating a single correlation coefficient. The two-stage
test procedure should offer protection against such distortions. In
addition, it was shown that greater latitude exists in terms of non-
distorting pooling when the subgroup mean differences are small, the
subgroup sample sizes are large, and the pooled within-subgroup corre-
lations are low to moderate. The calculated examples suggest limits
within which distortion does not seriously affect correlational results.
The types of subgroups to which this discussion refers are those resulting from natural dichotomies and those resulting from an arbitrary split where (a) the decision to split the total group was based on considerations other than that of ridding the data of non-linearity, and (b) middle range data has been discarded. When an arbitrary split is made to rid the total group data of non-linearity, the two subgroups may show evidence of different, but linear, relationships. According to the first stage test, if the two within-subgroup correlations are different, then it would appear unreasonable to even consider pooling the data on the basis of the original rationale for having made the split. On the other hand, if middle range data is not discarded when an arbitrary split is made for reasons other than ridding the data of non-linearity, then pooling would appear to be a reasonable step toward restoring the information contained within the total bivariate set of data.

Implications of this study relate to the use of the pooled within-subgroup correlation coefficient, to current practices in educational research, and to generalizations of this study in terms of pooling multiple subgroups in the calculation of correlation matrices. First, assuming no difference between the within-subgroup correlation coefficients (non-rejection of the first stage test), the pooled within-subgroup correlation coefficient is useful as a descriptive statistic with hypothesis-testing capabilities resulting from its equivalence to a first-order partial correlation coefficient. Second, in research involving multiple dependent measures that are analyzed via several ANOVA’s, rather than MANOVA, intercorrelations among the dependent measures should be assessed through pooled within-subgroup correlation coefficients rather than total group correlation coefficients. Otherwise,
the meaningfulness of the intercorrelations would be contingent upon the failure to find subgroup differences in all of the ANOVA's.

Last, considering the demonstrated varieties of possible distortion of a single correlation coefficient from the haphazard pooling of only two heterogeneous subgroups, the generalizations of these results must be inherently more complex, i.e., the effects of pooling multiple subgroups on correlation matrices. If, indeed, such complex distortion can be demonstrated, and it is desirable to pool data for reasons of parsimony, this suggests that further study should be devoted to the multivariate case and the development of appropriate test procedures.
References


Table 1

Values of $r_t$ Resulting from Three Patterns of Subgroup Centroid Differences and Five Total Group Sample Sizes: $r_T = .8000$

<table>
<thead>
<tr>
<th>Total group sample size</th>
<th>Significance levels for $t$ distribution when subgroup mean differences equal critical values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p &lt; .05$</td>
</tr>
<tr>
<td><strong>Pattern 1: $x_2 &gt; x_1$, $y_2 &gt; y_1$, both significant</strong></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.8799</td>
</tr>
<tr>
<td>50</td>
<td>.8155</td>
</tr>
<tr>
<td>100</td>
<td>.8077</td>
</tr>
<tr>
<td>200</td>
<td>.8039</td>
</tr>
<tr>
<td>1000</td>
<td>.8008</td>
</tr>
<tr>
<td><strong>Pattern 2: $x_2 &gt; x_1$, significant; $y_1 = y_2$</strong></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.6200</td>
</tr>
<tr>
<td>50</td>
<td>.7683</td>
</tr>
<tr>
<td>100</td>
<td>.7844</td>
</tr>
<tr>
<td>200</td>
<td>.7923</td>
</tr>
<tr>
<td>1000</td>
<td>.7983</td>
</tr>
<tr>
<td><strong>Pattern 3: $x_2 &gt; x_1$, $y_2 &lt; y_1$, both significant</strong></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.0813*</td>
</tr>
<tr>
<td>50</td>
<td>.6602*</td>
</tr>
<tr>
<td>100</td>
<td>.7305</td>
</tr>
<tr>
<td>200</td>
<td>.7653</td>
</tr>
<tr>
<td>1000</td>
<td>.7931</td>
</tr>
</tbody>
</table>

Note.--Asterisks refer to significance levels of unit-normal $z$ tests comparing $r_T$ and $r_w$.

* $p < .05$

** $p < .01$

*** $p < .001$

**** $p < .0001$
Table 2

Values of $r_t$ Resulting from Three Patterns of Subgroup Centroid Differences and Five Total Group Sample Sizes: $r_w = .4000$

<table>
<thead>
<tr>
<th>Total group sample size</th>
<th>Significance levels for $t$ distribution when subgroup mean differences equal critical values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p &lt; .05$</td>
</tr>
<tr>
<td><strong>Pattern 1: $X_2 &gt; X_1$, $Y_2 &gt; Y_1$, both significant</strong></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.6396</td>
</tr>
<tr>
<td>50</td>
<td>.4466</td>
</tr>
<tr>
<td>100</td>
<td>.4232</td>
</tr>
<tr>
<td>200</td>
<td>.4116</td>
</tr>
<tr>
<td>1000</td>
<td>.4023</td>
</tr>
<tr>
<td><strong>Pattern 2: $X_2 &gt; X_1$, significant; $Y_1 = Y_2$</strong></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.3100</td>
</tr>
<tr>
<td>50</td>
<td>.3842</td>
</tr>
<tr>
<td>100</td>
<td>.3922</td>
</tr>
<tr>
<td>200</td>
<td>.3961</td>
</tr>
<tr>
<td>1000</td>
<td>.3992</td>
</tr>
<tr>
<td><strong>Pattern 3: $X_2 &gt; X_1$, $Y_2 &lt; Y_1$, both significant</strong></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-.1590</td>
</tr>
<tr>
<td>50</td>
<td>.2913</td>
</tr>
<tr>
<td>100</td>
<td>.3459</td>
</tr>
<tr>
<td>200</td>
<td>.3730</td>
</tr>
<tr>
<td>1000</td>
<td>.3946</td>
</tr>
</tbody>
</table>

Note.—Asterisks refer to significance levels of unit-normal $z$ tests comparing $r_t$ and $r_w$:

- * $p < .05$
- ** $p < .01$
- *** $p < .001$
Table 3

Values of $r_t$ Resulting from One Pattern of Subgroup Centroid Differences and Five Total Group Sample Sizes: $r_w = 0.0000$

<table>
<thead>
<tr>
<th>Total group sample size</th>
<th>Significance levels for $t$ distribution when subgroup mean differences equal critical values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p&lt;.05$</td>
</tr>
<tr>
<td>Pattern 1: $X_2 &gt; X_1$, $Y_2 &gt; Y_1$, both significant</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.3993</td>
</tr>
<tr>
<td>50</td>
<td>.0777</td>
</tr>
<tr>
<td>100</td>
<td>.0386</td>
</tr>
<tr>
<td>200</td>
<td>.0193</td>
</tr>
<tr>
<td>1000</td>
<td>.0038</td>
</tr>
</tbody>
</table>

Note.--Asterisks refer to significance levels of unit-normal $z$ tests comparing $r_t$ and $r_w$.

* $p<.05$

** $p<.01$
FIG. 1. Nine exaggerated bivariate plots of total group data resulting from pooling heterogeneous subgroups.