Correcting Correlations When Predicting Success In College

Joe L. Saupe & Mardy T. Eimers
University of Missouri
Association for Institutional Research
Chicago, Illinois
May 29 – June 2, 2010
Introduction

• Low correlations between test scores and success in college

• Restricted population
  – do not apply for college
  – Apply/not accepted
  – Accepted/do not enroll
  – enroll part-time (and are excluded from analyses)
  – enroll full-time but do not complete the first year

• Most relevant population: all HS graduates
Introduction

• Predictors of success in college
  – indicators of success in HS
  – scores on admissions tests

• HS GPA and rank in class higher correlations than test scores

• Combination of test score and indicator of success in HS is the best predictor
Purpose of Study

• Illustrate techniques for correcting correlations for restriction in variance when predicting success in college

• Examine the correlations involving an admissions test score and indicators of success in HS with first-year college GPA

• Stimulate additional study
Formulas

• Classical measurement theory provided formulas for correcting correlations

• Variances are restricted due to selection on one or more variables

• Selecting college students on the basis of an admissions test or an indicator of HS success leads to a restriction in the variance of the selection variable
Formulas

• *Explicit* selection: selection on the basis of the predictor variable

• *Implicit* or *incidental* selection: selection based upon a variable that is related to the predictor variable

• Two of the three principal formulas are applicable for the present study

• Formulas require that the variance of a relevant variable in the unrestricted population be known
Case 1 - Incidental Selection on admissions test score; this formula not used

Case 2 – Explicit Selection: on admissions test score
- Selection based upon $X_1$ and the values of $s_1^2$, $s_2^2$, $S_1^2$, and $r_{12}$ are known.

$$R_{12} = \frac{S_1 r_{12}}{\sqrt{S_1^2 r_{12}^2 + s_1^2 - s_1^2 r_{12}^2}}$$

- Students are selected on the basis of a test score, $X_1$, and the variance of that score in the unrestricted population, $S_2^2$, is known.
Formulas

Case 3 – Explicit selection on a variable (admissions test score) related to predictor variable of interest (HS GPA)

- Selection on $X_3$ and $X_1$ is a third variable related to $X_3$. Here the values of $s_1^2$, $s_2^2$, $s_3^2$, $S_3^2$ $r_{12}$, $r_{13}$, and $r_{23}$ are known.

\[
R_{12} = \frac{r_{12} - r_{13}r_{23} + r_{13}r_{23}(S_3^2 / s_3^2)}{\sqrt{[1 - r_{23}^2 + r_{23}^2(S_3^2 / s_3^2)]}[1 - r_{12}^2 + r_{12}^2(S_3^2 / s_3^2)]}
\]

- Students are selected on the basis of a test score, $X_3$, the variance of that variable in the unrestricted population is known, and that variable is related to the predictor of interest, HS grade point average, $X_1$. 
The data

• First-time freshmen at a research university
  – full-time degree-seeking students
  – completed both semesters, and
  – complete data for the study variables

• Restricted to those with HS class percentile rank of 50 or greater in order to approximate the assumption of the correction formulas

• $N = 3,668$
The data

- **ACT-C** - ACT Composite Score
- **HSCPR** - HS Class Percentile Rank
- **NHSCPR** - Normalized HS Class Percentile
- **CCGPA** - HS Core Course GPA
- **FYGPA** - First-year College GPA
Descriptive statistics from restricted population

Table 1

<table>
<thead>
<tr>
<th>Statistic</th>
<th>ACT-C</th>
<th>NHSCPR</th>
<th>CCGPA</th>
<th>FYGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>25.65</td>
<td>60.13</td>
<td>3.50</td>
<td>3.05</td>
</tr>
<tr>
<td>s</td>
<td>3.71</td>
<td>6.32</td>
<td>0.38</td>
<td>0.67</td>
</tr>
<tr>
<td>s²</td>
<td>13.76</td>
<td>39.94</td>
<td>0.14</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Variable
Assumptions

- Linearity and homoscedasticity (but not normality)

- Students selected on the basis of HSCPR or NHSCPR (not strictly the case in study).

- Study population is restricted to students with HSCPR of 50 or greater in order to meet, to some degree, the assumption.
Assumptions

• Correction formulas require that the variance of a relevant variable in the unrestricted population, i.e., all HS graduates, be known

• The variance of NHSCPR for all HS graduates is $10^2$ or 100

• If the unrestricted population were all “college bound” HS graduates, the correlations in that population could be estimated using ACT-C and the variance of these scores in the population of ACT test-takers
Results

Table 2
Correlations

<table>
<thead>
<tr>
<th>Variable</th>
<th>NHSCPR</th>
<th>CCGPA</th>
<th>FYGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT-C</td>
<td>0.43</td>
<td>0.36</td>
<td>0.43</td>
</tr>
<tr>
<td>NHSCPR</td>
<td>0.78</td>
<td></td>
<td>0.49</td>
</tr>
<tr>
<td>CCGPA</td>
<td></td>
<td>0.56</td>
<td></td>
</tr>
</tbody>
</table>
Results

Correlation of ACT-C with FYGPA

• Use Case 3 formula with variables
  – $X_1 = \text{ACT-C}$, $X_2 = \text{FYGPA}$, and $X_3 = \text{NHSCR}$, and values from Table 1 and Table 2

• Result: $R_{12} = .56$
Results

Correlation of NHSCPR with FYGPA

• Use Case 2 formula with variables
  – \( X_1 = \text{NHSCPR} \) and \( X_2 = \text{FYGPA} \), and values from Table 1 and Table 2

• Result: \( R_{12} = .76 \)
Results

Correlation of CCGPA with FYGPA

• Use Case 3 formula with variables
  – \( X_1 = \text{CCGPA}, \ X_2 = \text{FYGPA}, \text{ and } X_3 = \text{NHSCPR}, \) and values from Table 1 and Table 2

• Result: \( R_{12} = .80 \)
Discussion

Table 3

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Restricted</th>
<th>Unrestricted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>Population</td>
</tr>
<tr>
<td>r</td>
<td>% of Var</td>
<td>R</td>
</tr>
<tr>
<td>ACT-C</td>
<td>0.43</td>
<td>18</td>
</tr>
<tr>
<td>NHSCPR</td>
<td>0.49</td>
<td>24</td>
</tr>
<tr>
<td>CCGPA</td>
<td>0.56</td>
<td>31</td>
</tr>
</tbody>
</table>

- Unrestricted population correlations (with FYGPA) and percent of variance values are estimates.
- Accuracy of the estimates is unknown, because they are affected by the violations of the assumption that students were elected solely on the basis of HSCPR.
Discussion

Four major findings:

1. Correlations increased appreciably from restricted to unrestricted populations: .43 to .56 for ACT-C, .49 to .76 for NHSCPR, and .56 to .80 for CCGPA.

2. Correlations of ACT-C with FYGPA, .43 to .56 are relatively modest, but the increase in percentage of variance predicted (78%) is not trivial.

3. Indicators of success in HS have higher correlations with FYGPA than admission test scores:
   - .49 for NHSCPR
   - .56 for CCGPA
   - .43 for ACT-C
Discussion

4. Increases in correlations from restricted to unrestricted populations are greater for indicators of HS success than for admissions test score, ACT-C:
   • from .43 to .56 (difference of .13) for ACT-C
   • from .49 to .76 (difference of .27) for NHSCPR
   • from .56 to .80 (difference of .24) for CCGPA

Pattern of increases in percentages of variance is similar
Further research

• Single institution study is not definitive
• Colleges and universities with different degrees of selectivity or which differ in other ways may yield different results
• Studies in which the assumption about the basis of selection is more closely met are needed
• Degree of sensitivity to violations of that assumption needs study.
Conclusions

• Restriction in range causes correlations calculated from enrolled student populations to understate true relationships between predictor and college success variables.

• Other variables, e.g., high school attended and unreliability in the correlated measures, also depress the correlations.

• Variables that predict college success are more accurate than what is generally shared in the literature and are more appropriate for use in admissions that is often argued.

• This study will be successful if it stimulates additional research.
Joe L. Saupe  
Professor Emeritus of Education  
Mardy T. Eimers  
Director, Institutional Research  
University of Missouri  
Columbia, Missouri 65211

Any questions please contact Mardy Eimers at 573-882-4077 or eimersm@missouri.edu