Addendum:

Factor Analysis of Explanatory Variables in an Achievement Production Function

James L. Phelps

Combining explanatory variables into factors instead of using individual variables in an achievement production function is advocated in several of the articles in this special issue. The following is a brief overview of factor analysis explaining and illustrating the reasoning for this technique. There is a linchpin: Factor analysis is an aspect of regression analysis which is used to estimate the relationships between an outcome and the explanatory variables of a production function.

This idea originated from the desire to find a single number—an index—representing a school's socioeconomic status (SES). The process started with a large number of possible explanatory variables and was reduced down to just those variables making a significant and consistent contribution to predicted achievement. The SES index became a part of a comprehensive achievement production function. The initial goal was easily accomplished via regression for any one year; however, there was a substantial difference in the statistically significant variables and the magnitude of their weightings across years. There was no logical justification for these differences. As it turned out, small differences in the correlation matrix across years produced large differences in results. What were the reasons? Was there a workable alternative addressing these vagaries?

Factor analysis searches for combinations of variables—the factors—based on the common variance among variables in a correlation matrix. When a factor or factors have been previously conceptualized as being associated, factor analysis can confirm the assumption and provide an estimate of the strength of the factor(s). In other words, confirmatory factor analysis determines if conceptually associated variables are statistically related. If factors have not been previously conceptualized as being related, exploratory factor analysis identifies combinations of variables which are statistically related—the factors—and provides information helpful for the conceptualization effort.

While different in purpose, factor analysis and regression analysis share similarities. Regression estimates the relationships between an outcome and several explanatory variables, taking into consideration the relationships among the explanatory variables. Factor analysis, in contrast, estimates the relationships only among combinations of explanatory variables. Step-wise regression first identifies the single explanatory variable extracting the maximum variance associated with an outcome variable, removes this variance, and then identifies the next variable extracting the maximum variance, and so on until all independent variables are exhausted. In contrast, factor analysis identifies a combination of explanatory variables extracting the maximum variance, removes this variance, and then identifies the next combination of variables extracting the maximum variance, and so on. Each factor is orthogonal; that is, it is uncorrelated, with no linear relationship to the others.

Factor analysis is frequently used to explore combinations of statistically related variables by setting the number of factors to be identified at a minimal number and working upwards. After all, the better explanations are usually the simplest explanations. After the factors, their constituent variables, and their weightings have been identified, the task remains to place the results into some coherent conceptual framework. Factor analysis does not do this; indeed, factor analysis can produce incoherent results when there is substantial collinearity among all the variables. On the other hand, if there is no correlation among the explanatory variables, each variable is a factor, an easily understood but infrequent occurrence. Factor analysis is valuable for investigating student achievement where most explanatory variables are correlated.

The principle of factor analysis is illustrated mathematically by the simplest case of regression between an achievement variable (correlation subscript 1) and two explanatory variables (subscripts 2 and 3). The amount of explained variance (R^2) is calculated by the formula:

$$R^2 = r^2_{12} + r^2_{13} - 2\, r_{12}r_{13}r_{23} / 1 - r^2_{23}$$

or

$$R^2 = (r^2_{12} / 1 - r^2_{23}) + (r^2_{13} / 1 - r^2_{23}) - (2\, r_{12}r_{13}r_{23} / 1 - r^2_{23})$$

If the correlation between the two explanatory variables is zero (r_{23}), the third term in the numerator is zero (and the denominator becomes1); hence no common variance exists, and the explained variance is the sum of the two squared correlations. In other words, each variable is a factor. In contrast, if the correlation between the two explanatory variables is greater than zero, the common variance is subtracted from the sum of the other variances. Because of the common variance, the two explanatory variables form a factor; that is, the two explanatory variables work cooperatively rather than independently to influence the outcome. The degree to which the variables work together is measured by the common variance. In stepwise regression, the explanatory variable with the largest correlation with the outcome variable is entered first, and the common variance subtraction is applied to the next variable entered, overestimating the influence of the first and underestimating the influence of the second. This explains why small differences in the correlation matrix produce large differences in regression results across years. The ambiguous interpretations of the common variance compound as more correlated explanatory variables are added into the regression equation. Moreover, there is a point where additional variables are no longer significant, and thus eliminated from consideration in the interpretation. Given this statistical reality, there is a workable alternative. The unique variance for each variable and the common variance among all explanatory variables can be combined into a factor predicated on an underlying theory explaining how the individual variables work together to achieve an outcome.

The notion of factors is incorporated into an achievement production function when socioeconomic status (SES) is included in a production function. Because there is no specific definition of SES, a combination of student and community characteristics is assembled as proxies to represent SES. The proxies are selected based on their conceptual logic, their statistical relationships among the variables, and their relationships with the outcome variable. In earlier papers, this notion of combining explanatory variables has also been applied to staff quantity with the variables of teachers, support teachers, teacher aides, and administrators, because these staffing roles work cooperatively to improve student achievement. Likewise, the variables of years experience, salary, age, and educational training are components of staff characteristics because these attributes combine to influence performance. Because of the substantial conceptual and statistical association of the variables within the concepts of staff quantity and staff characteristics, the use of factors seems logical. To further substantiate this position, these two conceptual factors—staff quantity and staff characteristics—are the foundation of confirmatory and exploratory factor analyses, addressing several questions. The examples are from a correlation matrix derived from the same data set described and used in the previous articles in this issue.

Are the proposed constituent explanatory variables related to the conceptual factor?

Tables 1 and 2 present the confirmatory factor analysis results for staff quantity and staff characteristics. The magnitude of association of the variables within the factor is measured in terms of factor loadings and amount of explained variance. The explained variance is calculated by dividing the squared factor loading by the number of explanatory variables. Only the relevant variables are included in the analysis. The factor analysis of staff quantity confirms the assumption that these staff roles are statistically associated. As might be expected, the contribution by teacher is highest, with administrators making little contribution to the explained variance. The factor analysis of staff characteristics confirms the assumption that these attributes are statistically associated. The contribution to the explained variance by graduate educational training (Masters Degree) is lower than other variables. Together, Tables 1 and 2 support the practice of combining explanatory variables into factors of staff quantity and staff characteristics for inclusion in an achievement production function.

Table 1
Factor Analysis of Staff Quantity

Variable	*Factor Loading*	*Squared*	*Percent*	*Variance*
Teacher	0.845	0.714	0.494	0.179
Administrator	0.099	0.010	0.007	0.002
Support	0.649	0.421	0.291	0.105
Aide	0.548	0.300	0.208	0.075
Sum		1.445		
Variance		0.361		0.361

When the constituent variables of both concepts are combined and analyzed, do they reasonably identify the two conceptual factors?

A separate exploratory factor analysis was conducted placing the constituent variables of both factors into a single analysis, restricted to two factors to determine if the analysis would identify the proposed factors. (See Table 3.) The analysis identified two factors, however, not the ones anticipated. Moreover, the resulting factors do not lead to a coherent explanation. Because of the collinearity of the variables, the staff characteristics overwhelmed the analysis, eliminating the staff quantity variables from consideration. This is an example of exploratory analysis where the factors do not lead to a coherent explanation.

Table 2
Factor Analysis of Staff Characteristics

Variable	*Factor Loading*	*Squared*	*Percent*	*Variance*
Years	0.767	0.588	0.274	0.147
Salary	0.755	0.570	0.265	0.143
Age	0.839	0.704	0.327	0.176
Masters Degree	0.537	0.288	0.134	0.072
Sum		2.151		
Variance		0.538		0.538

Table 3
Factor Analysis of Combined Explanatory Variables: Explained Variance of Contributing Variables

Variables	*Factor 1*	*Factor 2*
Staff Quantity		
Teacher	0.041	0.000
Administrator	0.014	0.0001
Support	0.001	0.006
Aide	0.032	0.002
Staff Characteristics		
Years	0.000	0.111
Salary	0.083	0.010
Age	0.002	0.110
Masters Degree	0.083	0.000
Sum	0.258	0.239

Table 4
Factor Analysis of Combined Explanatory Variables: Explained Variance

Variables	*Factor 1*	*Factor 2*	*Factor 3*
Staff Quantity			
Teacher	0.000	0.000	0.093
Administrator	0.000	0.025	0.000
Support	0.005	0.048	0.045
Aide	0.001	0.009	0.029
Staff Characteristics			
Years	0.111	0.000	0.000
Salary	0.015	0.079	0.010
Age	0.111	0.002	0.000
Masters Degree	0.001	0.056	0.027
Sum	0.244	0.220	0.205

When the constituent variables of both concepts are placed in the analysis, do they reasonably identify more than the two coherent factors?

An exploratory analysis was conducted on the same set of data allowing for three factors. (See Table 4.) Factor 1 incorporates years of service and age while the second factor incorporates support staff, salary, and masters degrees. The third combines teachers, support, and aides. Support is influential in both the second and third factor. All three factors are weaker in total variance than the ones previously identified. None of the factors reflect some higher-order concept. These results do not offer insights clearer than the analyses in Tables 1 and 2.

The first two examples confirm the statistical relationships among the component variables within the proposed staff quantity and staff characteristics factors. This occurs because the variables were preselected due to their logical association with the concept. In contrast, neatly formed factors do not emerge when all the variables, that are also correlated, are put into the analysis. Recall the three-variable regression formula: When explanatory variables are correlated, each explanatory variable cannot be a unique factor. This explains why regression results based on large numbers of correlated variables are most likely incoherent and conceptually unwise.

In these articles, the component variables are combined into regression factors and used to: (1) Report the standing of schools on the factors, rather than on individual variables; and (2) estimate the effectiveness of schools when these factors are statistically controlled. First, for each individual factor, the component variables are regressed against the achievement variable to obtain weightings, and these weightings are averaged over time.[1] The averaged weightings are then coefficients in an equation, representing the factor's relationship with the achievement variable. When the coefficients are entered into the equation for each school observation and evaluated, the results are a single number which best predicts the achievement. The result is an index combing the unique and common variance representing the standing for each school on each factor. This is done for SES, staff quantity, and staff characteristics. Now the achievement prediction equation has just three explanatory variables rather than a large number of variables.

Finally, the residuals of the yearly regression analysis are averaged to obtain an estimate of the school effectiveness. Averaging the residual is a common method in econometrics to estimate the fixed effect, i.e., the influence on achievement unique to each school. The details are included in this special issue.

In summary:

- Combining explanatory variables into factors for use in an achievement production function regression analysis is appropriate when the factor variables are conceptually and statistically related.
- Entering the individual explanatory variables separately into a production function regression analysis is appropriate only when the explanatory variables are conceptually independent and minimally correlated.
- Conversely, entering the individual explanatory variables separately into a production function regression analysis is problematic when the explanatory variables are conceptually related and substantially correlated.
- While helpful, factor analysis does not resolve all the issues inherent in regression analysis when a large number of variables are correlated. In these cases, a careful theoretical foundation is critical.

Throughout the special issue and this discussion, the purpose has been to link theory, evidence, and methodology to build a comprehensive and workable achievement production function. The underlying theory is based on what is generally accepted as being true: (1) Instructional staff work as a team to influence achievement; and (2) a combination of characteristics influence teacher behavior and performance. The evidence provided in Tables 1 and 2 supports the theory. Therefore, the logical method is to combine the variables identified conceptually and verified via factor analysis and use regression to obtain the weightings to construct an index for each factor. Finally, the indices representing the factors become the components of an achievement production function:[2]

Achievement = SES (9) + Staff Quantity (4) + Staff Characteristics (5) + Effectiveness

This comprehensive formulation brings a conceptual clarity, ease of explanation, coherence,[3] and simplicity not present when individual variables are the starting point of an achievement production function.[4]

Endnotes

[1] Because the weightings do not change over time, the best estimate of the true value is the average.

[2] The numbers in parentheses are the number of constituent variables in the factors.

[3] In an earlier effort, all the variables were entered into the equation, and it was virtually impossible to make a coherent explanation of the results because of the substantial correlation among the explanatory variables.

[4] With the variables included individually, there would be 18 mostly-correlated variables, with the dilemma of how to attribute the common variance and interpret the results.