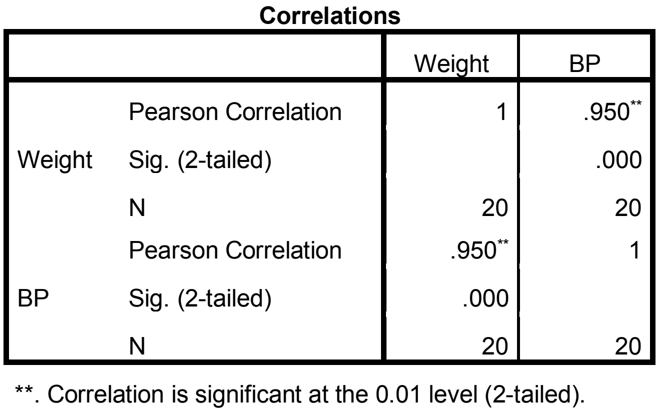
Multicollinearity

The term multicollinearity is used to refer to the extent to which independent variables are correlated. Multicollinearity exists when one independent variable is correlated with another independent variable, or if an independent variable is correlated with a linear combination of two or more independent variables.

It is always a good idea to check for collinearity in the data prior to analysis. A situation where it is likely to occur is, for example, when analyzing data containing respondents age and income, as these are bound to be highly correlated: as age goes up, income tends to increase. Another example, as shown below, is the relationship between weight and blood pressure:



Correlations

		Weight	BP
Weight	Pearson Correlation	1	.950**
	Sig. (2-tailed)		.000
	N	20	20
BP	Pearson Correlation	.950**	1
	Sig. (2-tailed)	.000	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

In this example, the correlation between these two variables is 0.950. Using both variables as predictors is likely to produce an error message in HLM indicating the presence of multicollinearity.

If the model

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is to be fitted, the solution generally would be of the form

$$b = \left(X^{'}X\right)^{-1} X^{'}Y$$

Obtaining the solution thus depends on obtaining the inverse of the matrix product $X^{'}X$. The matrix X represents the design matrix. If this product is singular (and do not have a unique inverse) an infinity of solutions exists.

Multicollinearity is a problem because a unique least-squares solution for regression coefficients is used as starting values in HLM. When two or more variables are perfectly correlated is present such a solution does not exist. In the case of highly correlated variables, it may still be problematical as the marginal contribution of that independent variable is influenced by other independent variables. Even when an analysis is possible, estimates for the regression coefficients may be unreliable and test of significance for them may be misleading.

In the context of a hierarchical linear model, the existence of high correlations between variables is a well-known cause of instability in the model. The correlations causing the problem may, however, be between two predictors at the same level, or it may be that a cross-level interaction is highly correlated either with the second-level variable or with the first-level variable, or both.

One way of dealing with this problem is by centering predictors entered into the model. In particular, centering of level-1 predictors around the respective group means may lower some of the correlations among the variables involved. When group mean centering is used, the correlations between second-level variables and both first-level variables and cross-level interactions are equal to zero, so that only the correlations between cross-level interactions and level-1 variables remain as a potential source of estimation problems.

The impact of high correlations on the numerical stability is also a function of the total amount of information in the actual data set used. Note, however, that centering impacts the interpretation of results and should be used with caution. It should be kept in mind that in HLM the intercept is represented in the design matrix by a column of 1's. If a predictor used in the level-1 model shows little or no variation, this variable will in effect be a duplicate of the intercept and the result would be a multicollinearity issue. The same problem would occur if data from, for example, single gender schools are analyzed in HLM, with the schools as the level-2 IDs. Should the gender of the students be introduced as predictor, it would be an exact duplicate of the intercept term within each level-2 unit.

In HLM, most reports by users concerning error messages noting collinearity/multicollinearity are caused by

- Near collinearity between a predictor with little or no variation and the intercept term, which is represented in the design matrix by a column of 1's.
- Fitting quadratic growth curves to a very short series of points so that a situation similar to that described above is the result.

The best place to look should HLM print an error message warning about collinearity / multicollinearity in the random part of the model is in the τ matrix/matrices given in the output file. Check all off-diagonal elements for correlations close to 1 or -1. Also check the diagonal elements of the τ matrix for any elements close to zero, as this may indicate that there is no indication of random variation in this slope.