Heywood cases represent the most common form of a series of related problems in confirmatory factor analysis and structural equation modeling. Other problems include factor loadings and factor correlations outside the usual range, large variances of parameter estimates, and high correlations between parameter estimates. The concept of empirical underidentification is used here to show how these problems can arise, and under what conditions they can be controlled. The discussion is centered around examples showing how small factor loadings, factor correlations near zero, and factor correlations near one can lead to empirical underidentification.

Structural Equation Models

Empirical Identification, Heywood Cases, and Related Problems

DAVID RINDSKOPF
City University of New York

Data analysts testing confirmatory factor analysis or structural equation models are often plagued by a variety of undesirable results: negative unique or residual variance estimates (Heywood cases), failure to converge to a solution, parameters that are outside reasonable limits, large standard errors of parameter estimates, and large correlations among parameter estimates. There have been few attempts to explain these problems or to offer guidance when they occur. One approach to the conceptualization of these problems was taken by van Driel (1978). His methods are not framed directly in terms of the natural parameters of the model, and are therefore not useful as guides to correcting the problems. McDonald (1982; McDonald and Krane, 1977, 1979) has also discussed some of these issues, in the context of identification, but has focused on the detection of nonidentification rather than its correction.

In this article, an alternative approach is taken based on the notion of empirical underidentification, which was discussed by

SOCIOLOGICAL METHODS & RESEARCH, Vol. 13 No. 1, August 1984 109-119

Kenny (1979). Kenny developed the idea of empirical underidentification and demonstrated it in several examples. This paper extends Kenny's work by developing a framework for analyzing causes of and solutions to problems in factor analysis and structural equation models, using empirical underidentification as the link between several problems encountered and the various causes of those problems.

The first part of the paper describes the concept of empirical underidentification, and shows how various situations common in factor analysis and structural modeling can cause it. It will be shown why empirical underidentification can result in the problems mentioned above. The examples will also show that the corrective action to take is not always obvious; for example, it is not always correct to remove a variable from an analysis when it has a negative error variance estimate, because the problem may be caused by another variable.

To understand the concept of empirical underidentification, one must first recognize that in establishing whether or not a model is identified, often certain seemingly innocuous assumptions are made that may not be true, and when they are not true, the model is not identified. For most models, then, one cannot say that the model is identified but only that it may be identified if certain conditions are true. This was noted by Koopmans (1949) and is implicit in Koopmans and Reiersol (1950), Reiersol (1950), and Anderson and Rubin (1956). Lawley and Maxwell (1971) discuss problems for identification in the one-factor, three-variable model resulting from various configurations of values of the population covariances. They did not discuss causes of the underlying problems, or strategies for detecting causes of problems in doing analyses.

The conditions for identification generally take the form of requiring that certain parameters not be zero or that parameters not equal one. The conditions that generally cause problems in factor analysis are factor loadings close to zero, factor correlations close to one, and factor correlations close to zero. In the following discussion, the terms "large," "small," "close to zero," and "close to one" are used without specifying exactly what they

mean. The exact values that might cause or prevent the problems discussed are rather fuzzy, and can change from one situation to the next. In addition, how large a standard error or correlation between parameter estimates must be before they are considered too large is a matter of judgment. In my judgment, a factor loading of less than .1 in absolute value should certainly be considered small, but one smaller than .2 in absolute value might be considered small under some circumstances. A correlation greater than .95 is "close to one," and a correlation greater than .90 often is. Similar remarks apply to correlations within .05 or .10 of zero. Further complications in making these judgments arise when the covariance matrix is analyzed instead of a correlation matrix, or in structural models when unstandardized parameter estimates are obtained.

The simples example of empirical underidentification can be seen in the following model specification:

$$\Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \qquad \Psi = \mathrm{Diag}(\psi_1, \psi_2, \psi_3) \qquad \Sigma = \Lambda\Lambda' + \Psi$$

This three-variable, one-factor model is generally thought to be exactly identified. As Kenny (1979), and McDonald and Krane (1979) have pointed out however, if one of the factor loadings is zero, the model is not identified. If there is a factor loading that is close to zero, some parameter estimates will have large standard errors. The standard errors that will be affected will not be the ones for the small factor loading, but rather the other factor loadings and error variances. Because a covariance matrix for three variables in which two covariances are almost zero is a very strong indication of which factor loading is close to zero, the standard error for that parameter will not be large. Given that terms involving that parameter appear in the denominator of the formulas for the other loadings, a value close to zero means that

the estimate of the loadings in which those values appear in the denominator will be unstable; that is, they will have large standard errors. As a result, the error variances for those variables will also have large standard errors, and the estimate of the error variance for one of them could be negative. In addition to these problems, notice that as one factor loading approaches zero, it is only the product of the other two loadings that is well determined. Therefore, the loadings could have any values, even outside the allowable range, as long as their product remains the same. They are, in other words, very highly (negatively) correlated, and when one estimate falls outside a certain range a Heywood case is automatically produced. This simple example shows how empirical underidentification can result in the whole range of problems described in the introduction.

The previous discussion was concerned with the population values of the covariances (and structural model parameters). More problems are possible when sample values are considered; a slight change from population values might make a negative error variance estimate possible. This follows as a consequence of the large standard errors of these parameters (for the population values, and therefore likely for the sample values). The usual judgment that is made in factor analysis is that when a variable is estimated to have a negative error variance, that variable may be in some sense faulty, and it is often eliminated from the model. In this example, it can be seen that another variable may be the cause of the negative error variance estimate. Although the situation as described so far is bad enough (from the standpoint of detecting the source of estimation problems), there are other potential sources of the same problems: violation of any of the assumptions of linearity, additivity, and normality might be the cause of patterns of observed correlations or covariances that, along with small factor loadings or intercorrelations or large factor intercorrelations, may be troublesome. Other sources of these problems are various forms of model misspecification such as omitted variables or factors, correlated errors, and omitted or extra causal paths or factor loadings.

In the first example, we examined the effect of a small factor loading. The next example shows the effect of a small correlation between factors. This example (which is also addressed by Kenny, 1979) is a two-factor, four-variable model in which two variables load only on the first factor, and the other two variables load only on the second factor. This model may be represented as:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ 0 & \lambda_3 \\ 0 & \lambda_4 \end{bmatrix} \qquad \Phi = \begin{bmatrix} 1 & \phi \\ \phi & 1 \end{bmatrix} \qquad \Psi = \text{Diag}(\psi_1, \psi_2, \psi_3, \psi_4)$$

$$\Sigma = \Lambda\Phi\Lambda' + \Psi$$

As Kenny notes, this model is identified only if the factors have a nonzero correlation. This can be seen intuitively by noting that if the factors do not correlate, the model is equivalent to two one-factor models, each with two observed measures, and neither of which is identified. One consequence of this is that, contrary to what is usually expected, one may make a model identified by removing, instead of adding, restrictions: a model in which the correlation is restricted to equal zero is not identified, but if the correlation is estimated (and is not close to zero) the model is identified.

For reasons similar to those described in the previous example, if the factor correlation is near zero (instead of exactly zero), the standard errors of many parameter estimates will be large. Also, sampling variability, nonlinearity, nonadditivity, and nonnormality might also produce covariance patterns in which standard errors are large.

The next example is one which has not been previously discussed in the literature on identification in factor analysis or structural models. The problem occurs when the correlation

between two factors is too high. A model for two factors and five variables that demonstrates this is:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & \lambda_4 \\ 0 & \lambda_5 \\ 0 & \lambda_6 \end{bmatrix} \qquad \Phi = \begin{bmatrix} 1 & \phi \\ \phi & 1 \end{bmatrix} \qquad \Psi = \text{Diag}(\psi_1, \psi_2, \psi_3, \psi_4, \psi_5)$$

Note that variable 3 is an indicator for both factors; this is the crucial point for empirical underidentification. As the correlation between the factors approaches one, the loadings for variable 3 become indeterminate because the two factors are almost identical. If the population values for both loadings were .5, then any combination of estimates which summed to one would result in an identical covariance matrix, as long as the error variance for variable 3 were appropriately adjusted. When we consider problems caused by sampling, nonlinearity, and so on, it becomes plausible to assume that for some observed covariance patterns the factor loadings might be rather extreme, such as 5 and -4, which would result in a large negative estimate for the error variance of that variable. Therefore, this is another possible cause of Heywood cases. The two factor loadings for variable 3 will have a large negative correlation and large standard errors; so again, a whole range of symptoms can result from this one problem.

MULTICOLLINEARITY AS EMPIRICAL UNDERIDENTIFICATION

One of the most persistently discussed problems in the area of multiple regression is that of multicollinearity, or high intercorrelations among independent variables. The results are regression

parameters with very large standard errors, so that the parameter estimates are very unstable. Multiple regression can be represented as a very simple path analysis model, which in turn is a special case of a structural equation model. Within this framework, the problem of multicollinearity can be viewed as another case of empirical underidentification.

There is an exact analogy with the previous factor analysis example, in which the factor correlation approached one. In the multiple regression case, an equation such as $Y = aX_1 + bX_2$, in which X_1 and X_2 correlate perfectly, cannot be solved exactly for a and b. If, for example, the population values for a and b are .5, then any combination of estimates for a and b that sum to 1 will give the same predicted values of Y as will the population values. When the correlation between X_1 and X_2 is much less than perfect, the parameters are identified. When the correlation is close to one, the standard errors will be large, and the parameters are ill-determined.

We might expect similar problems in the more general case of structural equation models when latent variables are highly correlated. This can be detected by examining these correlations directly, when available. A less complete check would be to determine that the correlations of the exogenous latent variables (Φ in LISREL notation) are smaller than one, and that the unique variances of the latent endogenous variables (diagonal elements of Ψ in LISREL models) are not too small. (Small values on the diagonal of Ψ do not indicate that some latent variables must be highly correlated, but only that they might be.)

OVERFACTORING

There are several conditions that all can be considered to be instances of overfactoring, or the inclusion of unnecessary factors. This occurs when a factor has (1) no large loadings, or (2) only one large loading, or (3) only two large loadings, and close to zero correlation with all other so-called real factors—that is, factors with two or more large loadings. In all of these cases, the

model will be empirically underidentified. In the first case, the factor can have correlations with other factors over a fairly wide range without substantially changing the estimated covariance matrix. In the second case, the one nonzero loading cannot be separated reliably from unique variance, and the result can be an impossibly large (much greater than 1) loading accompanied by a negative unique variance. In the third case, the lack of correlation between that factor and other factors has the same effect as in the two-factor, four-variable model considered previously. An implication of this is the well-known result of Anderson and Rubin (1956) that at least three measures of each factor are needed to identify orthogonal models.

Although a wide variety of models can result in the problems discussed here, the use of confirmatory factor models for multitrait multimethod matrix analysis seems to provide a large share of the uninterpretable solutions. Kenny (1979) reports such a case and describes empirical underidentification as the cause. Although Joreskog (1974) evidently did not encounter such problems in his analysis, the use of different start values for the data he analyzed can yield factor intercorrelations outside the (-1, 1) range. Boruch and Wolins (1970) report that 30 out of 40 analyses yielded results that were problematic. Many of the reported problem cases in the literature seem to be the result of overfactoring: some of the hypothesized factors do not exist. Although this would be expected most frequently for method factors, in Kenny's example it was a trait factor that caused the problem. In many cases, combining the factor that appears to be problematic with another factor with which it correlates highly can solve the problem. This was the case in the example reported in Joreskog (1974).

DISCUSSION AND RECOMMENDATIONS

The most widely used computer programs for covariance structure analysis are probably those written by Joreskog and his colleagues. The latest version of his LISREL program (Joreskog

and Sorbom, 1978; 1981) will attempt to determine whether or not a model is identified. The manual warns that this check is not foolproof and that, if in doubt, the user should check algebraically that the model is identified. My contention is that the program is probably right, and that when it declares a model unidentified that is algebraically identified, the most likely cause is empirical underidentification. When large standard errors are obtained for parameter estimates, empirical underidentification is the likely problem. Especially when dealing with a complicated model, a person trying to determine the identification status of a model algebraically may easily miss one or more conditions for empirical underidentification in the equations, but the computer program will always detect the problem. Of course, the program is determining whether there is empirical identification in the sample, not the population. If the sample size is reasonably large, this distinction is probably not important in terms of diagnosing problems.

Empirical underidentification is the mediating concept between a variety of causes, on the one hand, and effects, on the other. The causes of empirical underidentification can range from the problems with the model parameters described in detail above, to violation of assumptions of linearity, normality, and additivity, to omission of important factors, variables, or paths. Any of these might be the source of empirical underidentification, which then could result in one or more of the symptoms usually encountered by analysts: Parameter estimates outside possible bounds, large standard errors, large correlations among parameter estimates, and failure to converge.

The cause of and solution to some of the problems discussed here can be investigated more easily with the aid of a simple technique discussed by Rindskopf (1983). Although the LISREL program as usually implemented does not prevent Heywood cases, a simple trick in parameterizing factor analysis and structural equation models can effectively impose the inequality restrictions necessary to accomplish this. The technique is to use exogenous latent variables, whose variance is fixed at one, as unique

and residual variables, instead of parameterizing models in the usual way. In LISREL notation, this means that θ_δ, θ_ϵ, and Ψ are not used, but ξs whose variances are fixed at one are used in their place. This has the added effect of keeping the factor loadings in bounds, given that negative error variance estimates are often the result of an attempt to compensate for large factor loadings. The model will still be empirically underidentified with some large standard errors and highly correlated parameter estimates, but the results will be much easier to interpret and the model easier to change because the parameter estimates will make sense.

When an analyst encounters the problems that indicate possible empirical underidentification, the following strategy is suggested: First, check the data for signs of nonnormality and nonlinearity. Nonadditivity can sometimes be detected using multiple regression on a selection of the observed variables. If the program did not converge properly, and the parameter estimates look reasonable, then check the start values, which may have been too far from the solution (this is generally a problem only with a large model or very strange start values). If there are negative variance estimates, use the procedure described in Rindskopf (1983) and to which I refer above. This should also prevent impossible values of factor loadings, although there may still be problems with factor correlations. Next, look for signs of overfactoring: factors with no large loadings, or only one large loading, or two large loadings (if the factor has low correlations with other factors). Eliminate this factor or combine it with another factor (fixing its correlation with the other factor at one will accomplish this).

If the parameter estimates for some factor loadings are highly correlated, look for the cause in the variables that load on two or more of the factors. Then, either combine the factors, or do not allow any variable to load on more than one of the highly correlated set. Look also for factor correlations that are close to zero or path and structural coefficients that are close to zero. In the latter case, underidentification may or may not be the problem, and further investigation (algebraic solution of the identification equations) will probably be necessary.

REFERENCES

ANDERSON, T. W., and H. RUBIN (1956) "Statistical inference in factor analysis." Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 5: 111-150.

BORUCH, R. F., and L. WOLINS (1970) "A procedure for estimation of trait, method, and error variance attributable to a measure." Educ. and Psych. Measurement 30: 547-574.

JORESKOG, K. G. and D. SORBOM (1981) "LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods. Research Report 81-8 Uppsala, Sweden: University of Uppsala, Department of Statistics.

———(1978) LISREL IV: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: National Educational Resources.

KENNY, D. A. (1979) Correlation and causality. New York: Wiley.

KOOPMANS, T. C. (1949) "Identification problems in econometric model construction." Econometrica 17: 125-143.

———and O. REIERSOL (1950) "The identification of structural characteristics." Annals of Mathematical Statistics 21: 165-181.

LAWLEY, D. N., and A. E. MAXWELL (1971) Factor analysis as a statistical method. London: Butterworth.

McDONALD, R. P. (1982) "A note on the investigation of local and global identifiability." Psychometrika 47: 101-103.

———and W. R. KRANE (1979) "A Monte Carlo study of local identifiability and degrees of freedom in the asymptotic likelihood ratio test." British J. of Mathematical & Statistical Psychology 32: 121-132.

———(1977) "A note on local identifiability and degrees of freedom in the asymptotic likelihood ratio test." British J. of Mathematical and Statistical Psychology 30: 198-203.

REIERSOL, O. (1950) "On the identifiability of parameters in Thurstone's multiple factor analysis." Psychometrika 15: 121-149.

RINDSKOPF, D. (1983) "Parameterizing inequaltiy constraints on unique variances in linear structural models." Psychometrika 48: 73-83.

VAN DRIEL, O. P. (1978) "On various causes of improper solutions in maximum likelihood factor analysis." Psychometrika 43: 25-243.

David Rindskopf holds a joint appointment in the Education and Psychology Departments at the City University of New York Graduate Center. His areas of research interest include the methodological aspects of nonexperimental research, latent variable models, and qualitative data analysis.