Although the dictum “Correlation does not imply causation” is well rehearsed in research methods courses throughout the social sciences, its converse actually serves as the basis for path analysis and the larger structural equation modeling (SEM) family of quantitative data analysis methods (which also includes, for example, multiple regression analysis and confirmatory factor analysis). Specifically, under the right circumstances, “Causation does imply correlation.” That is to say, if one variable has a causal bearing on another, then a correlation (or covariance) should be observed in the data when all other relevant variables are controlled statistically and/or experimentally. Thus, path analysis may be described, in part, as the process of hypothesizing a model of causal (structural) relations among measured variables—as often represented in a path diagram—and then formally examining observed data relations for degree of fit with the initial hypotheses’ expected relations.

Most modern treatments of path analysis trace the technique’s beginnings to the work of the biometrician Sewell Wright (e.g., Wright, 1918), who first applied path analysis to correlations among bone measurements. The method was barely noticed in the social sciences until Otis Duncan (1966) and others introduced the technique in sociology. Spurred by methodological, computational, and applied developments mainly during the 1970s and 1980s by such scholars as K. G. Jöreskog, J. Ward Keesling, David Wiley, Dag Sörbom, Peter Bentler, and Bengt Muthén, countless applications have appeared throughout the social and behavioral sciences. For example, shifting focus from correlations to covariances allowed for the development of a formal statistical test of the fit between observed data and the hypothesized model. Furthermore, the possibility of including latent (i.e., unobserved) factors into path models and the development of more general estimation techniques (e.g., maximum likelihood, asymptotically distribution free) addressed major initial limitations of traditional correlational path analysis. Continuous refinements in specialized computer software throughout the 1980s, 1990s, and today (e.g., LISREL, EQS, AMOS, Mplus) have led to a simultaneous increase in technical sophistication and ease of use and thus have contributed to the proliferation of path analysis and more general SEM applications over the past three decades.

In this entry, typical steps in the path analysis process are introduced theoretically and via example. These steps include model conceptualization, parameter identification and estimation, effect decomposition, data-model fit assessment, and potential model modification and cross-validation. More detail about path analysis and SEM in general may be found in, for example, Bollen (1989), Kaplan (2000), Kline (1998), and Mueller (1996).

Foudational Principles

As an example, an education researcher might have a theory about the relations among four constructs: Mathematics Self-Concept (MSC), Reading Self-Concept (RSC), Task Goal Orientation (TGO),
weights. With regard to the two-headed arrow, as it represents a covariance between two variables, the symbol appearing atop the arrow is the familiar $\sigma$. Using these symbols, we may express the relations from the figure in two ways: as structural equations and as model-implied relations. Structural equations are regression-type equations expressing each endogenous variable as a function of its direct causal inputs. Assuming variables are mean centered (thereby eliminating the need for intercept terms), the structural equations for the current model are as follows:

$$V_3 = p_{31}V_1 + p_{32}V_2 + 1E_3,$$
$$V_4 = p_{41}V_1 + p_{43}V_3 + 1E_4.$$

These equations, along with any noncausal relations contained in the model (i.e., two-headed arrows), have implications for the variances and covariances one should observe in the data. Specifically, if the hypothesized model is true in the population, then the algebra of expected values applied to the structural equations yields the following model-implied variances (Var) and covariances (Cov) for the population:

$$\text{Var}(V_1) = \sigma_1^2$$
$$\text{Var}(V_2) = \sigma_2^2$$
$$\text{Var}(V_3) = p_{31}^2\sigma_1^2 + p_{32}^2\sigma_2^2 + 2p_{31}\sigma_1\sigma_2 + p_{32}\sigma_2 + \sigma_{E3}^2$$
$$\text{Var}(V_4) = p_{41}^2\sigma_1^2 + 2p_{43}p_{31}\sigma_1\sigma_2 + p_{41}\sigma_1 + 2p_{43}\sigma_2 + 2p_{31}\sigma_1p_{32} + \sigma_{E4}^2 + \sigma_{E3}^2$$
$$\text{Cov}(V_1, V_2) = \sigma_{21}$$
$$\text{Cov}(V_1, V_3) = p_{31}\sigma_1^2 + p_{32}\sigma_2$$
$$\text{Cov}(V_1, V_4) = p_{41}\sigma_1^2 + p_{31}p_{43}\sigma_1\sigma_2 + p_{32}p_{43}\sigma_2$$
$$\text{Cov}(V_2, V_3) = p_{32}\sigma_2^2 + p_{31}\sigma_2$$
$$\text{Cov}(V_2, V_4) = p_{32}p_{43}\sigma_2^2 + p_{41}\sigma_2$$
$$\text{Cov}(V_3, V_4) = p_{43}(p_{31}^2\sigma_1^2 + p_{32}^2\sigma_2^2 + 2p_{31}\sigma_1\sigma_2 + 2p_{32}\sigma_2 + \sigma_{E3}^2) + p_{31}p_{41}\sigma_1^2 + p_{32}p_{41}\sigma_2$$
These 10 equations constitute the covariance structure of the model, with 9 unknown population parameters appearing on the right side \( p_{31}, p_{32}, p_{41}, p_{43}, \sigma^2_{21}, \sigma^2_1, \sigma^2_2, \sigma^2_3, \sigma^2_4 \). If one had the population variances and covariances among the four variables, those numerical values could be placed on the left, and the unknowns on the right could be solved for (precisely, if the hypothesized model is correct). Given only sample (co)variances, such as those shown in Table 1 for a hypothetical sample of \( n = 1,000 \) ninth graders, they may similarly be placed on the left, and estimates of the parameters on the right may be derived. To this end, each parameter in a model must be expressible as a function of the (co)variances of the observed variables. When a system of such relations can be uniquely solved for the unknown parameters, the model is said to be just-identified. When multiple such expressions exist for one or more parameters, the model is overidentified; in this case, a best-fit (although not unique) estimate for each parameter is derived. If, however, at least one parameter cannot be expressed as a function of the observed variables’ (co)variances, the model is underidentified, and some or all parameters cannot be estimated on the basis of the data alone. This underidentification might be the result of the researcher attempting to impose a model that is too complex (i.e., too many parameters to be estimated) relative to the number of (co)variances of the observed variables. Fortunately, underidentification is rare in most path analysis applications, occurring only with less common model features (e.g., nonrecursive relations).

Table 1  Sample Matrix of Variances and Covariances

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mathematics self-concept)</td>
<td>1.951</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(reading self-concept)</td>
<td>-.308</td>
<td>1.623</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(task goal orientation)</td>
<td>.262</td>
<td>.242</td>
<td>1.781</td>
<td></td>
</tr>
<tr>
<td>V4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mathematics proficiency)</td>
<td>28.795</td>
<td>-4.317</td>
<td>17.091</td>
<td>1375.460</td>
</tr>
</tbody>
</table>

The model in Figure 1 is overidentified with 10 observed (co)variances and 9 parameters to be estimated: 1 covariance, 4 direct paths, 2 variable variances, and 2 error variances. The model thus has \( 10 - 9 = 1 \) degree of freedom \((df)\). Given that a model is just- or (preferably) overidentified, parameter estimates can be obtained through a variety of estimation methods. These include maximum likelihood and generalized least squares, both of which assume multivariate normality and are asymptotically equivalent as well as asymptotically distribution-free estimation methods that generally require a substantially larger sample size. These methods iteratively minimize a function of the discrepancy between the observed (co)variances and those reproduced by substitution of iteratively changing parameter estimates into the model-implied relations. Using any SEM software package, the maximum likelihood estimates for the key paths in the current model are found to be (in the variables’ original units) as follows: \( \hat{p}_{31} = .163, \hat{p}_{32} = .180, \hat{p}_{41} = 13.742, \hat{p}_{43} = 7.575, \) and \( \hat{\sigma}_2^2 = -.308 \) (all of which are statistically significant at a .05 level), Standardized path coefficients, similar to beta weights in multiple regression, would be .170, .172, .518, .273, and -.173, respectively. Before focusing on these main parameter estimates, however, we should consider whether there is any evidence suggesting data-model misfit and any statistical—and theoretically justifiable—rationale for modifying the hypothesized model.

DATA-MODEL FIT ASSESSMENT AND MODEL MODIFICATION

One of the advantages of modern path analysis is its ability to assess the quality of the fit between the data and the model. A multitude of measures exists that assist the researcher in deciding whether to reject or tentatively retain an a priori specified overidentified model. In general, measures to assess the fit between the (co)variances observed in the data and those reproduced by the model and its parameter estimates can be classified into three categories: absolute, parsimonious, and incremental. Absolute fit indices are those that improve as the discrepancy between observed and reproduced (co)variances decreases, that is, as the number of parameters approaches the number of nonredundant observed (co)variances. Examples of such measures include the model chi-square statistic that tests the stringent null hypothesis that there is no data-model misfit in the population, the standardized root mean square residual (SRMR) that roughly assesses the average standardized discrepancy between observed and reproduced (co)variances, and the goodness-of-fit index (GFI) designed to evaluate the
amount of observed (co)variance information that can be accounted for by the model.

Parsimonious fit indices take into account not just the overall absolute fit but also the degree of model complexity (i.e., number of parameters) required to achieve that fit. Indices such as the adjusted goodness-of-fit index (AGFI), the parsimonious goodness-of-fit index (PGFI), and the root mean square error of approximation (RMSEA) indicate greatest data-model fit when data have reasonable absolute fit and models are relatively simple (i.e., few parameters). Finally, incremental fit indices such as the normed fit index (NFI) and the comparative fit index (CFI) gauge the data-model fit of a hypothesized model relative to that of a baseline model with fewer parameters.

The three types of fit indices together help the researcher to converge on a decision regarding a path model’s acceptability. Hu and Bentler (1999) suggested joint criteria for accepting a model, such as CFI values of 0.96 or greater together with SRMR values less than 0.09 (or with RMSEA values less than 0.06). In the current example, the nonsignificant model chi-square of 2.831 (with 1 df) indicates that the observed (co)variances could reasonably occur if our model correctly depicted the true population relations. Although bearing good news for the current model, this statistic is typically ignored as it is notoriously sensitive to very small and theoretically trivial model misspecifications under large sample conditions. As such, other fit indices are generally preferred for model evaluation. For the current model, no appreciable data-model inconsistency is suggested when applying the standards presented above: CFI = 0.997, SRMR = 0.013, and RMSEA = 0.043.

After the data-model fit has been assessed, a decision about that model’s worth must be reached. Acceptable fit indices usually lead to the conclusion that no present evidence exists warranting a rejection of the model or the theory underlying it. This is not to say that the model and theory have been confirmed, much less proven as correct; rather, the current path model remains as one of possibly many that satisfactorily explain the relations among the observed variables. On the other hand, when fit indices suggest a potential data-model misfit, one might be reluctant to dismiss the model entirely. Instead, attempts are often made to modify the model post hoc, through the addition of paths, so that acceptable fit indices can be obtained. Most path analysis software packages will facilitate such model “improvement” by providing modification indices (Lagrange multiplier tests) indicating what changes in the model could reap the greatest increase in absolute fit, that is, decrease in the model chi-square statistic. Although such indices constitute a potentially useful tool for remedying incorrectly specified models, it seems imperative to warn against an atheoretical hunt for the model with the best fit. Many alternative models exist that can explain the observed data equally well; hence, attempted modifications must be based on a sound understanding of the specific theory underlying the model. Furthermore, when modifications and reanalyses of the data are based solely on data-model misfit information, subsequent fit results might be due largely to chance rather than true improvements to the model. Modified structures therefore should be cross-validated with an independent sample whenever possible. From the current analysis, none of the modification indices suggested changes to the model that would result in a statistically significant improvement in data-model fit (i.e., a significant decrease in the model chi-square statistic); indeed, as the overall model chi-square value is 2.831 (significance level > 0.05), there is no room for statistically significant improvement. Thus, no model modification information is reported here.

At last, given satisfactory data-model fit, one may draw conclusions regarding the specific model relations. Interpretation of path coefficients is similar to that of regression coefficients but generally with a causal bent. For example, $\hat{p}_{31} = 0.163$ implies that, under the hypothesized model, a 1-unit increase in scores on the MSC scale directly causes a 0.163-unit increase in scores on the TGO scale, on average, holding all else constant. Recall that this was a statistically significant impact. Alternatively, one may interpret the standardized path coefficients, such as the corresponding value of 0.170. This implies that, under the hypothesized model, a 1 standard deviation increase in scores on the MSC scale directly causes a 0.170 standard deviation increase in scores on the TGO scale, on average, holding all else constant. Other direct paths, unstandardized and standardized, are interpreted similarly as the direct effect of one variable on another under the hypothesized model.

In addition to the direct effects, other effect coefficients may be derived as well. Consider the modeled relation between RSC and MP. RSC does not have a direct effect on MP, but it does have a direct effect on TGO, which, in turn, has a direct effect on MP. Thus. RSC is said to have an indirect effect on MP.
because, operationally speaking, one may follow a
series of two or more single-headed arrows flowing in
the same direction to get from one variable to the other.
The magnitude of the indirect effect is the product of
its constituent direct effects, (0.180)(7.575) \approx 1.364
for the unstandardized solution. This unstandardized
indirect effect value implies that a 1-unit increase in
scores on the RSC scale indirectly causes a 1.364-unit
increase in scores on the MP test, on average, holding
all else constant. Similarly, for the standardized solu-
tion, the standardized indirect effect is (0.172)(0.273)
\approx 0.047, which implies that a 1 standard deviation
increase in scores on the RSC scale indirectly causes a
.047 standard deviation increase in scores on the MP
test, on average, holding all else constant.

Finally, notice that MSC actually has both a direct
effect and an indirect effect on MP. In addition to
the unstandardized and standardized direct effects of
13.742 and 0.518, respectively, the indirect effect may
be computed as (0.163)(7.575) \approx 1.235 in the unstan-
dardized metric and (0.170)(0.273) = 0.046 in the
standardized metric. This implies that the total causal
impact of MSC on MP, the total effect, is the sum of
the direct and indirect effects. For the unstandardized
solution, this is 13.742 + 1.235 \approx 14.977, implying
that a 1-unit increase in scores on the MSC scale causes
in total a 14.977-unit increase in scores on the MP test,
on average, holding all else constant. Similarly, for the
standardized solution, this is 0.518 + 0.046 = 0.564,
implying that a 1 standard deviation increase in scores
on the MSC scale causes in total a 0.564 standard
development increase in scores on the MP test, on aver-
age, holding all else constant. Thus, through this type of
effect decomposition, which uses the full context
of the hypothesized model, one can learn much more
about the relations among the variables than would be
provided by the typically atheoretical correlations or
covariances alone.

SUMMARY

Path analysis has become established as an impor-
tant analysis tool for many areas of the social and
behavioral sciences. It belongs to the family of struc-
tural equation modeling techniques that allow for the
investigation of causal relations among observed and
latent variables in a priori specified, theory-derived
models. The main advantage of path analysis lies in
its ability to aid researchers in bridging the often-
observed gap between theory and observation. As has
been highlighted in this entry, path analysis is best
understood as a process starting with the theoretical
and proceeding to the statistical. If the theoretical
underpinnings are ill-conceived, then interpretability
of all that follows statistically can become compro-
mised. Researchers interested in more detail regarding
path analysis, as well as its extension to the larger
family of SEM methods, are referred to such resources
as Bollen (1989), Kaplan (2000), Kline (1998), and
Mueller (1996). Parallel methods involving categori-
cal variables, which have foundations in LOG-LINEAR
MODELING, may also be of interest (see Goodman,

—Gregory R. Hancock and Ralph O. Mueller

REFERENCES

New York: John Wiley.
American Journal of Sociology, 72, 1–16.
contingency tables when some variables are posterior to
others: A modified path analysis approach. Biometrika, 60,
179–192.
indexes in covariance structure analysis: Conventional crite-
ria versus new alternatives. Structural Equation Modeling:
equation modeling. New York: Guilford.
modeling: An introduction to LISREL and EQS. New York:
Springer.
Wright, S. (1918). On the nature of size factors. Genetics,
3, 367–374.