Methodologies for Conducting Research on Giftedness

EDITED BY
Bruce Thompson and Rena F. Subotnik

American Psychological Association
Washington, DC
PROMISE AND PITFALLS OF STRUCTURAL EQUATION MODELING IN GIFTED RESEARCH

REX B. KLINE

The family of statistical techniques that make up structural equation modeling (SEM) offers many potential advantages in education research, including studies about gifted students. These techniques are highly versatile and permit the evaluation of a wide range of hypotheses, including those about direct or indirect effects, measurement, or mean differences on either observed or latent variables. They have also become quite popular among researchers. Indeed, it is increasingly difficult to look through an issue of an education research journal and not find at least one article in which results of SEM analyses are reported. In gifted research, SEM has been used less often, but there are more and more such studies in this area, too. However, there are some potential pitfalls of using SEM in gifted research that are inherent to the study of a special population, including range restriction, regression effects, and the need for large samples. Accordingly, the goals of this chapter are to (a) review the general characteristics of SEM with special consideration of possible advantages in educational research and (b) consider specific potential pitfalls of using SEM in gifted research.
CRITICAL ISSUES IN SEM

Outlined next are important issues in basically all applications of SEM. Later sections deal with problems specific to the use of SEM in education research in general and in gifted research in particular.

Terms and Hypotheses

Other terms for SEM include covariance structure analysis and covariance structure modeling, but both of these labels are restrictive. This is because although covariances are analyzed in all applications of SEM, it is also possible to analyze means (and mean differences), too. That is, all models estimated in SEM have a covariance structure, but depending on the data and hypotheses, the model may have an optional mean structure, too. Most structural equation models described in the literature do not involve the analysis of means, but the possibility to analyze both means and covariances gives SEM much flexibility.

A covariance structure concerns a specific pattern of relations, or covariances, among a set of observed or latent variables. For observed variables X and Y, the sample covariance is calculated as the product $r_{XY} SD_X SD_Y$. In SEM, the researcher specifies a statistical model of the covariance structure that he or she presumes reflects the population structure. This structure often includes latent variables that correspond to hypothetical constructs and the observed variables that are presumed to measure those constructs. Observed variables in SEM can be categorical, ordinal, or continuous (i.e., quantitative), but all latent variables analyzed in SEM are continuous. There are other statistical techniques for analyzing categorical latent variables, which are also referred to as classes, but classes are not analyzed in SEM. The covariance structure specified by the researcher provides a "map," or a set of hypotheses, about how the observed and latent variables should covary. All applications of SEM involve the analysis of covariance structures.

It is also possible in SEM to analyze means. This is accomplished through the specification of a mean structure in addition to a covariance structure. A mean structure is basically a set of instructions for the computer to estimate the means of variables included in that structure. These variables may include observed variables or latent variables that correspond to hypothetical constructs. Other statistical techniques, such as the analysis of variance (ANOVA), analyze means of observed variables but not also of latent variables. Although means are not analyzed in most SEM studies, the possibility to do so lends much flexibility to the analysis. For example, sometimes we want to test hypotheses about hypothetical constructs (i.e., latent variables) by examining relations (i.e., usually covariances) among the observed variables, but we also want to test whether estimated means on these latent vari-
ables are equal across constructs or on a given construct across different
groups. That is, the hypotheses to be tested concern both covariances and
means. At other times, we are not interested in means on the latent variables
and focus only on trying to discern (interpret) latent variables, based on
investigation of the covariances among the measured variables. For example,
in one case we might only want to know, "How many factors or latent vari-
ables underlie the responses of gifted students to IQ test items?" In another
instance, we might want to know both, such as "How many factors or latent
variables underlie the responses of gifted students to IQ test items?" and "Do
boys and girls have different means on each of the given IQ factors?"

**Specification and Respecification**

Perhaps the single most important step in SEM occurs at the very begin-
ning when the researcher specifies the model to be analyzed. This specifica-
tion includes information about things such as which variables are presumed
to affect other variables (i.e., directionalities of effects) and also about the
correspondence between observed and latent variables (e.g., measurement).
These a priori specifications reflect the researcher's hypotheses, and together
they comprise the model to be evaluated. Model specification is the single
most important step in SEM. This is because the scientific merit of all that
follows depends much on the ability of the researcher to specify a reasonably
accurate model, one that stands at least some chance of being true in the pop-
ulation. That is, the adage "garbage in, garbage out" applies to SEM—and to
any other statistical modeling technique.

The requirement that the researcher specify an entire model before the
analysis can begin makes SEM more confirmatory than exploratory. But as
often happens in SEM, the data may not be consistent with the model, which
means that analysis typically takes a more exploratory turn as revised models
are tested with the same data. This is respecification, also referred to as a spec-
ification search. Just as in more conventional statistical methods, respecifica-
tion should be guided much more by substantive considerations than by
purely statistical ones. For example, stepwise methods in multiple regression
select predictors for entry into or removal from the equation based solely on
the statistical significance of regression coefficients. However, stepwise regres-
sion is generally unacceptable due to extreme capitalization on chance (i.e.,
the results are unlikely to replicate; see Thompson, 1995).

The parallel to stepwise regression in SEM concerns the use of modifica-
tion indexes, which are tests of whether adding a particular parameter to the
model would result in a statistically significant improvement in model fit.
Some SEM computer programs even offer an "automatic modification" option
where parameters are automatically added to the model if their associated
modification indexes are statistically significant. However, such procedures represent extreme capitalization on chance, just as do stepwise methods regression. Also, the results of some computer simulation studies indicate that respecification based solely on modification indexes is unlikely to lead to discovery of the true model (e.g., Silvia & MacCallum, 1988). Instead, respecification should be guided by the researcher's knowledge of extant theory and empirical results (i.e., base specification and respecification on the same grounds).

Note that the cautions about respecification just mentioned presume the analysis of a model in a single data set. These cautions may not apply when the researcher has a very large data set that is randomly split into two or more subsamples, an empirically based specification search is conducted in the first subsample, and the fit of the resulting model in the other subsamples is evaluated. This strategy is known as internal replication, and results that replicate across two or more subsamples may be more stable than ones found in a single data set. Internal replication in SEM would require a very large sample size, and probably for this reason there are relatively few examples of internal replication in the SEM literature. Even rarer still is external replication, where a structural equation model specified by one researcher is tested in a different sample collected by a different researcher in a different setting. External replication is a gold standard in science, but, unfortunately, it is rare in SEM.

Sample Size

Large samples are generally required in SEM. Although there have been attempts to develop ways to apply SEM in smaller samples (e.g., Marsh & Hau, 1999), such methods are not widely used and can be more difficult to apply. Problems in the analysis are more likely to occur, parameter estimates may not be very precise, and the power of statistical tests may be low, all due to the absence of a large sample size. In general, a sample size of less than 100 cases may be too small to analyze even a very basic model. Between 150 and 200 cases is a better minimum, but perhaps even 200 cases is an insufficient amount to analyze more complex models, which require the estimation of more parameters. A suggestion for models with latent variables that correspond to constructs (i.e., they have a measurement model) is to think about minimum sample size relative to model complexity. Ideally, one should have at least 20 cases for every estimated parameter (i.e., a 20:1 ratio); a 10:1 ratio is less ideal but may suffice (Jackson, 2003).

Data Screening

It is necessary to thoroughly "clean and screen" the data before analyzing them in SEM. This is because data-related problems can make such com-
puter tools yield an illogical solution or even crash. It is also true that the default estimation method in most SEM computer tools, maximum likelihood (ML), requires certain assumptions about the distributional characteristics of the data. One is multivariate normality for outcome (dependent) variables. This requirement is critical and cannot be violated with impunity; otherwise, the values of parameter estimates or statistical indexes of model fit may be inaccurate. If obvious skew or kurtosis is apparent, then corrective action, such as applying transformations in order to normalize the scores, is necessary when using ML estimation. There are other, more specialized estimation methods for nonnormal distributions, but they may be more difficult to apply or require even larger sample sizes.

Another critical issue is that of missing data. Not all SEM computer programs can process raw data files where some cases have missing observations. Thus, it may be necessary to address this problem first in a general program for statistics before conducting SEM. But just how missing data are handled in SEM can adversely affect the results. For example, the use of pairwise deletion of incomplete cases can result in a data matrix, such as a covariance matrix, that cannot be analyzed. This happens when the matrix contains values that would be mathematically impossible to obtain in a sample with no missing data. Unfortunately, there are no free lunches concerning data integrity in SEM—see Kline (2005, chap. 3) for more information. Along the same lines, McCoach (2003) noted that although "structural equation models may look impressive, SEM cannot salvage studies that contain badly measured constructs, inappropriate samples, or faulty designs" (p. 43).

Ease of Use Should Not Mean Suspension of Judgment

Virtually all computer tools for SEM allow data and model specification with a series of equations written in the native syntax of each individual program. Of course, specification in syntax requires knowledge of that syntax. As an alternative, some SEM computer tools, such as Amos, EQS, LISREL, and mxGraph, also permit the user to specify the model by drawing it on the screen. The user is typically provided with a palette of drawing tools in a graphical editor for drawing the model. Also, some drawing editors prevent the user from making "illegal" specifications with the drawing tools. The capability to draw a model on the screen helps beginners to be productive right away because it is not necessary to learn program syntax.

A possible drawback is that such ease of use could encourage the application of SEM in uninformed or careless ways. Steiger (2000) made the related point that the emphasis on ease of use in advertisements for some computer tools (e.g., SEM made easy!) can give beginners the false impression that SEM itself is easy. However, nothing could be further from the truth. Things
can and do go wrong in SEM analyses, and this is more likely to happen when analyzing more complex models. When (not if) problems crop up, a researcher will need all of his or her expertise about the data and model in order to "help" the computer with the analysis. This is in part why Pedhazur and Schmelkin (1991) advised concerning statistical analyses in general that "no amount of technical proficiency will do you any good, if you do not think" (p. 2).

Proper Reporting

There are some serious shortcomings in the way that results of SEM analyses are often reported. For example, MacCallum and Austin (2000) reviewed 500 SEM analyses published in 16 different research journals. In about 50% of these analyses, the reporting of parameter estimates was incomplete. The most common mistake was to report the standardized solution only. This is a problem because the default estimation method in most SEM computer programs, maximum likelihood, calculates standard errors for unstandardized estimates only. These standard errors are the denominators of statistical tests of individual parameter estimates. Many researchers do not realize that results of these tests may not apply in the standardized solution. This means that if an unstandardized estimate is statistically significant, then its standardized counterpart is not guaranteed to be statistically significant, too. There is a special method known as constrained estimation (optimization) that can calculate correct standard errors for standardized estimates, but it is not available in all SEM computer programs.

In about 25% of the published studies reviewed by MacCallum and Austin (2000), it was not specified whether the raw data, a covariance matrix, or a correlation matrix was analyzed. Probably most researchers input raw data files, but another option is to submit a matrix summary of the data. If a raw data file is submitted, an SEM computer program will create its own matrix, which is then analyzed. If means are not analyzed, a covariance matrix—or a correlation matrix and the standard deviations—is generally required when submitting a matrix summary. This is because ML estimation assumes that the variables are unstandardized. If a correlation matrix is analyzed instead of a covariance matrix, then the results—including values of standard errors and model fit statistics—may be wrong. The method of constrained estimation can be used to correctly analyze a correlation matrix.

In about 10% of the studies reviewed by MacCallum and Austin (2000), the model specified was not described in enough detail to replicate the analysis. An example of this kind of problem is the failure to describe the association between observed and latent variables. All of the shortcomings just described are serious. They are also surprising considering that there are published guidelines for the reporting of results in SEM (e.g., McDonald & Ho, 2002).
Assessing Model Fit

A related issue concerns the reporting of values of statistical indexes of model fit. Literally dozens of alternative fit statistics exist, and the output of SEM computer programs often lists the values of many more such statistics than a researcher would ever report. This embarrassment of statistical riches means that it can be difficult to decide which subset of fit indexes to report. It also increases the temptation for selective reporting—that is, for the researcher to report only those fit indexes with favorable values. Elsewhere I recommended that a minimal set of fit indexes to report include the model chi-square statistic with its degrees of freedom, the Steiger-Lind root mean square error of approximation (RMSEA) with its 90% confidence interval, the Bentler comparative fit index (CFI), and the standardized root mean squared residual (SRMR; Kline, 2005, pp. 133-145).

The fit statistics just mentioned measure somewhat different aspects of model fit. Briefly, the RMSEA measures the degree of approximate fit of the model taking account of sample size and model complexity, and values less than .05 may indicate good approximate fit. Another favorable result is when the value of the upper limit of the 90% confidence interval for the RMSEA is less than .10. The CFI ranges from 0 to 1.00, where 1.00 indicates perfect fit. Its value indicates the relative improvement in fit of the researcher’s model over a baseline model that usually assumes zero covariances between all pairs of observed variables. Values of the CFI greater than .90 suggest good fit. The SRMR is a standardized summary of the difference between sample and model-implied covariances, and values less than .10 are considered favorable. Note that even higher (e.g., CFI) or lower (e.g., RMSEA, SRMR) values are expected by many researchers, and over time, informed by Monte Carlo research, the trend has been to set increasingly demanding cutoffs for describing model fit as adequate (e.g., Fan, Thompson, & Wang, 1999; Hu & Bentler, 1999).

It is unacceptable to report the value of a single fit index. This is because no single fit statistic captures all aspects of model fit. Also, the guidelines just suggested for cutting values of fit statistics that indicate good fit of the model to the data (e.g., RMSEA < .05) are just that. This means that such cutting values should not be blindly applied without considering other aspects of the model, such as whether the individual parameter estimates make theoretical sense or whether most differences between observed and model-implied covariances or correlations are in fact small. As noted by Robert and Pashler (2000), good statistical fit of a model indicates little about (a) theory flexibility (e.g., what it cannot explain), (b) variability of the data (e.g., whether the data can rule out what the theory cannot explain), and (c) the likelihoods of other outcomes. These are all matters of rational, not statistical, analysis.
Consider Equivalent Models

For most structural equation models, it is possible to generate at least one equivalent version that fits the data just as well but with a different configuration of effects among the same variables. Some of these equivalent versions may include effects in the reverse direction compared with the original model. For example, if variable $Y_1$ is presumed to affect $Y_2$ in the original model (i.e., $Y_1 \rightarrow Y_2$), there may be an equivalent version in which the direction of that effect is reversed (i.e., $Y_2 \rightarrow Y_1$), but both models will have equal fit to the data. Specifically, equivalent models have equal values of fit statistics, including the model chi-square, the RMSEA, and the other indexes described earlier. For any structural equation model, there may be many equivalent versions; thus, the researcher is obliged to explain why his or her preferred model is superior over mathematically equivalent ones. Equivalent versions of structural equation models are generated by applying rule described in Kline (2005, pp. 153–156, 192–194, 229); see also Hershberger (2006). Unfortunately, it seems that authors of articles in which SEM results are reported rarely acknowledge the existence of equivalent models (MacCallum & Austin, 2000).

Causality It Ain't

Pardon the poor diction, but sometimes a little everyday language, plainly spoken, gets the point across. And the point is that certain regression coefficients in structural equation models that estimate directional effects can be interpreted as indicating causality only if the model is correctly specified in the first place. That is, if the causal structure is known in advance, then SEM could be used to estimate the magnitudes of the direct effects implied by that structure. However, this is certainly not how most researchers use SEM. Instead, they specify a hypothetical model of causality, and then they evaluate how well that model fits the data. If a satisfactory model is found after respecification, then about all that can be concluded is that the data are consistent with the causal structure represented in the model. This is especially true in nonexperimental studies with no design elements that directly support causal inference, such as random assignment, control groups, and the measurement of presumed causes before presumed effects. This is in part why Wilkinson and APA Task Force on Statistical Inference (1999) emphasized that the use of SEM computer programs “rarely yields any results that have any interpretation as causal effects” (p. 600).

It also seems that, in general, too many education researchers are not as cautious as they should be when inferring causality from correlational data. For example, Robinson, Levin, Thomas, Pituch, and Vaughn (2007) reviewed...
EXHIBIT 7.1
of Structural Equation Modeling

1. Do not estimate models in small samples.
2. Do not standardize variables or analyze a correlation except when using a special method for standardized variables; otherwise, analyze the raw data or a covariance matrix.
3. Given roughly equal fit, more parsimonious models are preferred over more complex models.
4. Verify distributional assumptions, such as multivariate normality, of the estimation method used.
5. Apply rational considerations—including practical and theoretical significance—when evaluating model adequacy.
6. Use multiple statistical criteria (fit indexes) to evaluate different aspects of the correspondence between the model and the data.
7. In structural regression models, evaluate the measurement model by itself before estimating the whole model.
8. Test plausible alternative models, when they exist.
9. Do not respecify a model without a priori theoretical justification.
10. Remember that infinitely many models—including equivalent models based on the same variables and numbers of paths—can fit the same data.


About 275 articles published in five different journals in the area of teaching and learning. They found that (a) the proportion of studies based on experimental or quasi-experimental designs (i.e., with researcher-manipulated variables) declined from 45% in 1994 to 33% in 2004. Nevertheless, (b) the proportion of nonexperimental studies containing claims for causality increased from 34% in 1994 to 43% in 2004. Robinson et al. suggested that an increased but uncritical use of SEM in education research may be a factor behind these results.

Listed in Exhibit 7.1 is a summary of Thompson's (2000) "10 commandments" of SEM. Each point deals with a critical issue in SEM, and most were mentioned earlier. However, it helps to see them all in one place. See also Kline (2005, pp. 313–324) for a discussion of 44 different ways to fool yourself with SEM and McCoach, Black, and O'Connell (2007) for an outline of common misinterpretations of SEM in school psychology research.

TYPES OF SEM ANALYSES WITH EDUCATION RESEARCH EXAMPLES

Described next are major types of structural equation models with specific examples from the education research literature. These examples illustrate some of the potential promise of various types of analyses in SEM.
Path Models

There are two characteristics of a path model: (a) there is a single measure (indicator) of each hypothetical variable (construct), and (b) the only latent variables are the error (residual) terms of outcome variables. The latter are called endogenous variables in SEM, and these variables in path models are represented as the consequences of other variables. The error term for an endogenous variable is called a disturbance, which represents all omitted causes of the corresponding endogenous variable. In contrast, exogenous variables are represented as having effects on other variables, but whatever causes exogenous variables is not represented in a path model. It is assumed in path analysis that every exogenous variable is measured without error (i.e., there is perfect score reliability). This requirement is unrealistic in practice. Measurement error in endogenous variables is reflected in their disturbances, but this implies that estimation of the proportion of unexplained variance is confounded with that of measurement error.

For an analysis described by Cramond, Matthews-Morgan, and Bandalos (2005), an IQ test and the Figural form of the Torrance Tests of Creative Thinking (TTCT; Torrance, 1966) were administered in 1959 to a sample of high school students. The version of the TTCT studied by Cramond et al. generated four scale scores, including Fluency, Flexibility, Originality, and Elaboration, but the IQ test yielded a single score. The lifetime creative achievements of the participants 40 years later were classified by expert raters in terms of quality and quantity.

A path model of the 40-year predictive validity of IQ and the TTAC is presented in Figure 7.1(a). The TTAC is represented in this model with a single total score across the four scale scores; likewise, creative achievement is represented by the sum of the quality and quantity scores. Observed variables are represented in diagrams of path models by squares or rectangles and latent variables by circles or ellipses. The IQ variable is exogenous because its cause is not represented in the model. In contrast, the TTAC and creative achievement variables are specified as endogenous. Every endogenous variable has a disturbance, which is also considered exogenous. The two-headed curved arrows that exit and reenter the same variable in the figure represent the variances of the exogenous variables, which are free to vary. The numbers (1) that appear in the figure are scaling constants for latent variables. In Figure 7.1(a), general intelligence and creative thinking are each specified as direct causes of later creative achievement. Creative thinking is also specified as a mediator of the effect of general intelligence on creative achievement. A mediator variable is part of an indirect effect, and here it reflects the hypothesis that creative thinking "passes on" some of the prior effect of general intelligence to creative achievement.
Figure 7.1. (a) A path model of the predictive validity of IQ and creative thinking. (b) A structural regression model for the same basic research question. CAT = creative achievement total; D = disturbance; E = error; TTCT = Torrance Tests of Creative Thinking (Torrance, 1966).
Structural Regression Models

An alternative kind of model similar to the one actually analyzed by Cramond et al. (2005) is presented in Figure 7.1(b). This is a structural regression (SR) model, which has both a structural part and a measurement part. As with path models, SR models permit the specification of presumed direct and indirect effects. Specifically, the SR model in Figure 7.1(b) contains the same basic structural relations among the general intelligence, creative thinking, and creative achievement variables as the path model in Figure 7.1(a). However, these effects in SR models can involve latent variables represented as measured by multiple indicators. For example, the four scale scores of the TTAC are specified as the indicators of a creative thinking factor in Figure 7.1(b), and the quantity and quality variables are represented as the indicators of a creative achievement factor. The paths from the factors to the indicators are direct effects, and here they reflect the presumed influence of the factors on their respective indicators. In contrast, there is only one measure of general intelligence (IQ scores), so no explicit measurement model for this indicator is represented in the SR model of Figure 7.1(b). This specification reflects the reality that there are times when a researcher has only one measure of some construct.

Each indicator in the measurement part of the model in Figure 7.1(b) has its own error term, which reflects systematic variance not explained by the corresponding factor and measurement error. Also, each of two endogenous factors in the structural part of the model has its own disturbance. This combination in an SR model means that the proportion of unexplained variance in endogenous factors (e.g., creative thinking) is estimated controlling for measurement error in its indicators (e.g., $E_{\text{Int}}, E_{\text{Fut}}, E_{\text{Creat}}, E_{\text{H}]})$. However, because there is no measurement error term for the single indicator of general intelligence in Figure 7.1(b), it is assumed that IQ scores are perfectly reliable, which is just as unrealistic as for the path model in Figure 7.1(a). Cramond et al. (2005) reported acceptable fit of this basic SR model to the data. They also found that general intelligence explained about 9% of the variance in creative thinking and that IQ and creative thinking together explained about 50% of the variance in creative achievement. Considering the time span in this longitudinal study (40 years), the latter result is impressive.

Confirmatory Factor Analysis Models

When analyzing an SR model, it is often important to first verify its measurement model—that is, to test the hypotheses about the correspondence between the indicators and the factors they are supposed to measure. If these hypotheses are wrong, then knowing relations among the factors
specified in the structural part of the model may be of little value (Thompson, 2000). For example, if the two-factor measurement model implied by the SR model in Figure 7.1(b) did not explain covariance patterns among the six indicators, then the fit of the whole SR model may be poor and the path coefficients may have little interpretive value. This is why many researchers use a two-step method to analyze SR models described by Anderson and Gerbing (1988), as follows: In the first step, researchers evaluate the measurement model implied by the original SR model. If this model is rejected, then it must be respecified. Given a satisfactory measurement model, the second step involves the testing of hypotheses about direct and indirect effects; that is, the structural part of the SR model is now analyzed. In this way, effects of misspecification of the measurement model are isolated from those of the structural model.

Confirmatory factor analysis (CFA) is the SEM technique for evaluating pure measurement models. Of all SEM techniques, it is probably CFA that has been applied most often in education research, especially in studies of construct validity. There are three characteristics of standard CFA models: (a) Each indicator is represented as having two causes, a single underlying factor it is supposed to measure and all other sources of variation represented by its error term. (b) These measurement errors are independent of each other. (c) Finally, all associations between every pair of factors are unanalyzed. That is, there are no direct effects between factors in CFA models, but such effects are a part of SR models (e.g., Figure 7.1(a)). Presented in Figure 7.2(a) is a standard CFA measurement model for the Mental Processing scale of the first edition Kaufman Assessment Battery for Children (KABC-I) (Kaufman & Kaufman, 1983), an individually administered cognitive ability test for children ages 2½ to 12½ years old. The test's authors claimed that the eight subtests represented in Figure 7.2(a) measure two factors, sequential processing and simultaneous processing. The three tasks believed to reflect sequential processing all require the correct recall of auditory or visual stimuli in a particular order, but more holistic, less order dependent reasoning is required for the five tasks thought to reflect simultaneous processing. The curved line with two arrowheads that connects the two factors in Figure 7.2(a) represents the covariance between them. This symbol also designates an unanalyzed association between two exogenous variables. If this symbol were replaced with a direct effect (e.g., Sequential → Simultaneous), then the whole model would be an SR model, not a CFA model.

The results of several CFA analyses of the KABC-I conducted in the 1980s and 1990s—including some with gifted students (Cameron et al., 1997)—generally supported the two-factor model presented in Figure 7.2(a). However, other results indicated that some subtests, such as Hand Movements, may actually measure both factors and that some of the measurement errors
Figure 7.2. (a) Two-factor CFA measurement model of the first edition of the Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983). (b) Latent growth model of change in level of externalization over 3 years. $E =$ error.

may covary (e.g., Kline, 2005, pp. 185-189). It may be possible to include both types of effects just mentioned in CFA measurement models. For example, the assumption that an indicator is multidimensional in that it reflects more than one domain corresponds to the specification that it loads on more one factor. The specification of error covariances, or unanalyzed associations between pairs of error terms, reflects the assumption that the two corresponding indicators share something in common that is not explicitly represented in the model. This "something" could be a common method of measurement, such as self-report, or autocorrelated error in the case of repeated measures variables. See Reynolds, Keith, Fine, Fisher, and Low (2007) for results of CFA analyses about the construct validity of the second edition of the Kaufman Assessment Battery for Children (KABC-II; Kaufman & Kaufman, 2004).

It is crucial to not commit the naming fallacy: Just because a factor is named does not mean that it is understood or even correctly labeled. Factors require some sort of designation, however, if for no reason other than com-
ommunication of the results, especially in a diagram. Although verbal labels are clearer than more abstract symbols (e.g., $A, \xi$), they should be viewed as conveniences and not as explanations. For example, just because a two-factor model of the KABC-I may fit the data does not prove that one factor is sequential processing and the other factor is simultaneous processing (see Figure 7.2a). Based on their CFA results, Keith and Dunbar (1984) suggested that an alternative label for the “sequential processing” factor of the KABC-I is “short-term memory,” and for the “simultaneous processing” factor, an alternative label is “nonverbal reasoning.” Which among a set of alternative labels for the same factor is correct is a matter of rational analysis.

Latent Growth Models

If a sample is tested on at least three occasions and all cases are tested at the same intervals, then it may be possible to analyze in SEM a latent growth model (LGM). Besides its covariance structure, an LGM has a mean structure, too. Presented in Figure 7.2(b) is an example of an LGM similar to one analyzed by Deković, Buist, and Reitz (2004). In a longitudinal study, Deković et al. collected annual measures of externalization over a 3-year period in a sample of adolescents. The mean structure in the figure is represented in part by the graphic symbol $\Delta$, which stands for a constant that equals 1.0 for every case. In SEM, this constant is treated as an exogenous variable (even though its variance is zero) that has direct or indirect effects on other variables in the model except for residual terms. The unstandardized coefficients for effects of the constant are interpreted as either means or intercepts. This is the same basic method carried out “behind the scenes” when a modern computer program for regression calculates an intercept.

In the covariance structure of the LGM in Figure 7.2(b), each annual measurement of externalization is represented as an indicator of two latent growth factors, Initial Status (IS) and Linear Change (LC). The IS factor represents the baseline level of externalization, adjusted for measurement error. Because the IS factor is analogous to an intercept in a regression equation, the unstandardized loadings of all indicators on this factor are fixed to 1. In contrast, loadings on the LC factor are fixed to constants that correspond to the times of measurement, beginning with zero for the first measurement and ending with 2 for the last. Because these constants (0, 1, 2) are positive and evenly spaced, they specify a positive linear trend, but one that is adjusted for measurement error when the model is estimated. The IS and LC factors are specified to covary, and the estimate of this covariance indicates the degree to which initial levels of externalization predict rates of subsequent linear change, again corrected for measurement error. A positive estimated covariance would indicate that adolescents with higher initial levels of externalization show
greater rates of linear increase over time, and a negative estimated covariance would indicate just the opposite.

There are no error covariances in the LGM of Figure 7.2(b). In models with at least four measurement occasions, it may be possible to specify that the measurement errors of adjacent assessments are assumed to covary, among other possibilities. Error covariances reflect dependencies among scores from the same cases across different measurement occasions.

In their sample, Deković et al. (2004) found that the levels of externalization increased in a positive linear way from year to year, but levels of initial externalization were not related to the rate of linear change. That is, adolescents with lower versus higher levels of externalization at the first observation did not differ appreciably in their subsequent rates of change. In follow-up analyses, Deković et al. added to the model three predictors of latent growth factors—adolescent gender, quality of parent relationships, and quality of peer relationships. All three predictors together explained about 30% of variance in the initial status factor but only about 3% of the variance in linear change factor. The single best predictor of initial level of externalization was quality of parent relationships, but it did not tell much about the rate of change over time in behavior problems.

Multiple-Sample SEM

Essentially any type of structural equation model can be analyzed across multiple samples. The main question addressed in a multiple-sample SEM is whether values of estimated model parameters vary appreciably across the groups. If so, then (a) group membership moderates the relations specified in the model (i.e., there is a group × model interaction), and (b) separate estimates of some model parameters may be needed for each group. There are many examples in the education research literature of multiple-sample SEM. For instance, Willett and Sayer (1996) analyzed latent growth models of change in reading and arithmetic at three ages (7, 11, and 16 years) across samples of children who differed in health status (healthy, chronic asthma, seizure disorder). They found that healthy children and those with chronic asthma had generally similar growth curves for academic skills, but children with a seizure disorder showed declining trajectories over time, a result that could be due to side effects of anticonvulsant medication. In a series of multiple-sample confirmatory factor analyses, Reynolds et al. (2007) concluded that the KABC-II seems to measure the same basic constructs across the age range 3 to 18 years. In multiple-sample SEM, it is possible to evaluate models across basically any grouping variable, including age, gender, diagnosis, placement in regular versus special education, and so on.
POTENTIAL PITFALLS OF SEM IN GIFTED RESEARCH

Three different potential problems of using SEM in gifted research are described next. One is the requirement for large sample sizes. Because gifted students make up a relatively small proportion of all schoolchildren (1% to 5%), it can be difficult for researchers in this area to collect samples large enough for proper analysis in SEM. The sample size problem is an issue in the study of special populations in general, so it is not unique to gifted research. Also, there are now published reports, two of which are described later in this chapter, of the use of SEM in large samples of gifted students, so the sample size problem is not insurmountable. Two other serious challenges are considered next—range restriction and regression artifacts.

Range Restriction

Gifted students are selected because their scores on measures including classroom grades, nomination reports, verbal or nonverbal cognitive ability scales, or standardized achievement tests are relatively high. There is ongoing debate about optimal ways to apply such measures in order to identify and select gifted students, especially in minority or underrepresented populations; for example, see the special issue of the journal *Theory Into Practice* (Ford & Moore, 2005). However, there is no doubt that variability on many cognitive or scholastic variables is restricted within samples of gifted students. Assuming linear associations, absolute values of bivariate correlations tend to be reduced through range restriction on either variable, and the amount of the difference between population and restricted sample correlations can be substantial. The use of measures that generate unreliable scores can also truncate sample correlations. Because covariances are made up of correlations and variances, their values for cognitive and scholastic variables may be restricted in gifted samples, too. Covariances are the basic statistic analyzed in SEM, so effects of range restriction could affect the results when models are tested with data from gifted students. An example follows.

Suppose that the XYZ test is a measure of general cognitive ability for students in elementary school. It is standardized for use across the whole range of student ability. The subtests of the XYZ test are placed on four different scales according to a theoretical measurement model. Reported in the test's manual are results of confirmatory factor analyses conducted within the standardization sample, and these results support the measurement model. A researcher administers the XYZ test within a large sample of gifted students. He or she specifies and tests the same model in a CFA but finds that the model's fit to the data is poor. Inspection of the parameter estimates indicates that the estimated factor correlations are much lower and that indicator error
variances are much higher than the corresponding values reported in the test manual. Based on these results, the researcher concludes that the XYZ test does not measure the same cognitive abilities among gifted students as it does in the general population. This conclusion may be unsound, however, and the reason is that this pattern of results is expected due to range restriction. In this case, it would safer to conclude that the construct validity of the XYZ tests holds across the whole range of ability, given the results reported in the test's manual.

There are various statistical methods that correct sample statistics for effects of range restriction, but they generally assume that (a) the unrestricted population variance is known and (b) the researcher wishes to generalize the corrected results to the unrestricted population (Sackett & Yang, 2000). The former requirement is not generally a problem with standardized tests, for which the population variance may be known. For example, the standardized score metric $\mu = 100.0, \sigma = 15.0$ (i.e., $\sigma^2 = 225.0$) is widely used for cognitive ability or achievement tests. However, if there is no interest in generalizing from a restricted sample to an unrestricted population, then statistical corrections for range restriction are of little use. In SEM analyses of cognitive or achievement variables measured in gifted samples, the possibility that the results are the consequences of range restriction is quite strong. Range restriction would present less of a problem when other kinds of variables are analyzed, specifically those weakly correlated with IQ or grades. These kinds of variables include self-esteem, personality traits, general psychological adjustment, family characteristics, and so on. Ranges of individual differences among gifted students on such variables may not differ substantially from those expected in the general population of students.

**Regression Artifacts**

This issue concerns the phenomenon of statistical regression to the mean, which refers to the tendency for cases with extreme scores to obtain less extreme scores when retested on the same or similar measures. Regression artifacts are a concern whenever cases are selected because they had scores either lower or higher than average and in studies where cases with extreme scores are matched across different conditions. Regression to the mean is caused by imperfect correlations (i.e., $|r| < 1.00$) between two sets of scores, and anything that lowers the correlation, such as score unreliability, increases it.

Regression effects should be expected in gifted research due to how gifted students are selected (i.e., because they have extreme high scores). Lohman (2006) reminded us that about half of students with scores in the top 3% of the distribution in one year will not obtain scores in the top 3% in the
next year due to regression to the mean, and this is true even for very reliable scores. If we forget about regression effects, then we could fool ourselves into thinking that many, or perhaps most, students identified as gifted in one year are no longer so the next. However, this result is exactly what we expect given that test scores basically never have perfect predictive validity, especially over one year. Regression effects can be reduced by measuring status on selection variables more than once and then averaging the scores together before classifying students as gifted or not. This reduces error in selection in part because effects of individual outlier scores tend to cancel out when scores are summed. It may also detect students who attain high scores on selection variables at a later point in time.

Although latent variables in structural equation models are continuous variables only, it is possible to test models with categorical latent variables in other types of statistical techniques. Using such techniques, it would be silly to test a model with predictors of whether students' classification as gifted or not remains stable over time when such stability is not expected even when working with very reliable scores, given regression effects. In SEM, it would be possible to specify and test a structural model of the predictive validity of variables used to select gifted students, but effects due to range restriction may compromise the results (i.e., underestimation of predictive power is expected).

SELECTED EXAMPLES OF SEM IN GIFTED RESEARCH

Considered next are two recent examples of SEM analyses conducted in large samples of gifted or high-ability students. An SR model with both measurement and structural components was analyzed in the first example, and a path model was estimated across separate samples by gender in the second example. These examples illustrate some of the kinds of hypotheses that can be evaluated in SEM when just covariance structures are estimated (i.e., means were not analyzed in these studies) in gifted research.

Within a sample of 624 gifted students in Hong Kong between the ages of 9 and 19 years (age $M = 13.0$ years, $SD = 2.3$ years; 49% boys, 51% girls), Chan (2006) administered measures of emotional intelligence (EI), social coping (SC), and psychological distress (PD). Variances on psychological or social variables among gifted students may be less affected by range restriction compared with cognitive or achievement variables. This may be true in part because, as noted by Chan, high ability does not make one immune from emotional, or social, or other kinds of personal problems. Two types of emotional intelligence were each assessed with two indicators, including self-relevant EI (self-management, utilization) and other-relevant EI (empathy, social skills). Likewise, two separate factors of SC were each assessed by two indicators,
including avoidant SC (denial, avoidance) and self-interaction SC (helping others, peer acceptance). The PD factor was measured by a total of five indicators (health concern, sleep problems, anxiety, dysphoria, suicidal thoughts). In a two-step analysis, Chan first verified that the measurement model just described had acceptable fit to the data before testing alternative SR models. In one SR model, effects of EI on PD were represented as entirely indirect through SC (i.e., a mediation-effect model). In another, direct effects of EI on PD were added (i.e., a mediation-direct-effect model). The latter model, although more complex than the former model, did not fit the data appreciably better than the simpler model. That is, the final results were consistent with the view that enhancement of EI alone may be insufficient to enhance psychological well-being or reduce distress among gifted students. Instead, an approach that recognizes the need to enhance the EI of gifted students relevant to SC by reducing avoidant coping and increasing empathy and social skills is consistent with the final mediation-effect model. Chan acknowledged that (a) these results only failed to disconfirm the final mediation-effect model, and (b) there are competitive models that could be found to be equally viable given similar testing methods.

A path model tested by Speirs Neumeister and Finch (2006) represented the hypothesis that parenting style affects the quality of a sense of attachment among children, which in turn predicts the degree of two different kinds of perfectionism. One is self-oriented perfectionism, where individuals impose high standards on themselves and evaluate their performance against these standards. The other type of perfectionism is socially prescribed, where individuals believe that high standards are established by others and that they are obliged to meet these standards. Speirs Neumeister and Finch speculated that self-oriented perfectionism would be associated more with motivation for mastery or performance approach compared with socially prescribed perfectionism, which would be associated more with a fear of failure due to the belief that self-worth depends on meeting the high standards of others. In a sample of 265 first-year university honor students (125 men, 140 women; total SAT score $M = 1396.1$), Speirs Neumeister and Finch administered separate measures of the parenting styles of mothers versus fathers, quality of attachment, self-oriented perfectionism and socially prescribed perfectionism, and achievement motivation (approach, avoidance, mastery). In a multiple-sample SEM, a path model with direct effects from parenting style variables to attachment, from attachment to perfectionism, and then from perfectionism to motivation was tested across samples of female versus male students. The results indicated that less-supportive parenting styles predicted less-secure attachment, but the effect of mother's style was stronger than that of father's style. Also, the general relation just described was even stronger for female students than for male students. For both women and men, it was found that (a) less-secure

166  REX B. KLINE
attachment predicted higher levels of perfectionism, and (b) the relation of perfectionism to achievement motivation was consistent with the hypotheses described earlier. This issue of equivalent or alternative path models was not addressed by the authors.

CONCLUSION

The price of the great diversity of research hypotheses that can be evaluated in SEM is that these techniques make several requirements, including the need to carefully specify the model in the first place, select measures that generate reliable scores, screen the data and check distributional assumptions before analyzing them, base model respecification on substantive grounds, and properly and accurately interpret and report the results. In education research, SEM is already widely used, but as noted, there are some serious shortcomings with its application in too many studies. Probably due to the need for large sample sizes, SEM has been used less often in gifted research, but its flexibility offers researchers in this area many potential advantages. However, these same researchers should be prepared to address the potential pitfalls of range restriction and regression effects, depending on the variables included in the model and the hypotheses tested about gifted students.

REFERENCES


