Building a Better Model: An Introduction to Structural Equation Modelling

David L Streiner, PhD[1]

Confirmatory factor analysis (CFA) and structural equation modelling (SEM) are powerful extensions of path analysis, which was described in a previous article in this series. CFA differs from the more traditional exploratory factor analysis in that the relations among the variables are specified a priori, which permits more powerful tests of construct validity for scales. It can also be used to compare different versions of a scale (for example, English and French) and to determine whether the scale performs equivalently in different groups (for example, men and women). SEM expands on path analysis by allowing paths to be drawn between latent variables (which, in other techniques, are called factors or hypothetical constructs), that is, variables that are not seen directly but, rather, through their effect on observable variables, such as questionnaires and behavioural measures. Each latent variable and its associated measured variables form small CFAs, with the added advantage that the correlations among the variables can be corrected for the unreliability of the measures.

(Can J Psychiatry 2006;51:317–324)

Information on funding and support and author affiliations appears at the end of the article.

Highlights

- CFA is an extension of exploratory factor analysis that allows for more powerful tests of the construct validity of a scale and the comparison of the equivalence of the scale across different versions and different populations.
- SEM extends path analysis in that relations among latent variables can be examined.

Key Words: *structural equation modelling, path analysis, confirmatory factor analysis*

In a previous article in this series, I discussed a powerful analytic technique called path analysis (1). Very briefly, path analysis is an extension of multiple regression that allows us to consider more than one DV at a time and, more important, allows variables to be both DVs and IVs. In other words, it permits us to consider chains of association, such that variable A can influence variable B, and B in turn can affect C. To avoid confusion about what to call variable B—it is a DV because it is affected by A, but it is also an IV because it is a predictor of C—we avoid those terms entirely. Instead, we substitute the terms exogenous variables for those that aren't influenced by any other variable in the model and endogenous variables for those that are (and this was supposed to reduce confusion?).

However, one limitation of path analysis is that it can handle only variables that are observed. At first glance, this hardly seems like a limitation: after all, if we can't observe a variable, we surely can't measure and analyze it. It has the flavour of the "ether" that supposedly permeated space but had the unique properties that it couldn't be seen, tasted, felt, or perceived in any other way—which led scientists down the garden path for centuries.

In fact, we deal with unobservable variables all the time, although we use other terms for them. In personality theory and test construction, they are called hypothetical constructs; in factor analysis, they are referred to as factors; and in SEM, the technique we will be considering here, they are known as latent variables. So much for the use

of consistent terminology to explain exactly what we mean. Whatever they're called, though, they refer to the same thing—variables that we cannot observe directly but know about through their purported effects on phenomena we can observe. This would apply to concepts such as intelligence, anxiety, depression, quality of life, coping style, schizophrenia, locus of control, and hundreds of others we encounter every day in psychiatry and psychology. Let's use anxiety as a model, recognizing that the principles apply equally well to the other constructs.

According to one theory (2), anxiety has 4 facets—cognitive, affective, behavioural, and physiological—that are themselves unobservable hypothetical constructs. When we say that a person is anxious, what we mean is that he or she is showing observable behaviours that are manifestations of one or more of these facets. In the physiological realm, for instance, there may be tachycardia, shortness of breath, and sweatiness (all of which are measurable); in the cognitive realm, there may be decreased ability to concentrate and hypersensitivity to perceived threats (again measurable). We postulate that these tend to occur together because they are all produced by the anxiety. In other words, anxiety is something we hypothesize to tie together observable phenomena that are correlated with each other to some degree. In a similar way, we do not see schizophrenia or intelligence but only a constellation of observed behaviours that tend to occur together and that we postulate are caused by some underlying mechanism.

SEM is an extension of path analysis that allows us to examine the relations among both measured and latent variables. (To add to the confusion, SEM is also used as an abbreviation for the standard error of the mean and the standard error of measurement; however, its meaning is usually clear from the context.) It does this by combining path analysis with a form of factor analysis called CFA, so it is probably easiest to begin with a discussion of CFA and how it differs from the more commonly encountered forms of factor analysis. (A reminder for those from the Maritime provinces of Canada: CFA does not stand for "come from away," or visitors from the rest of Canada, as one anonymous reviewer suggested.)

Abbreviations used in this article

ANX	anxiety
CFA	confirmatory factor analysis
DV	dependent variable
EFA	exploratory factor analysis
HSM	high school math scores
IV	independent variable
PNP	photonumerophobia: the fear that our fear of numbers will come to light
QOL	quality of life
RMSEA	root mean square error of approximation
SEM	structural equation modelling
SMR	squared multiple correlation
TAX	tax return errors

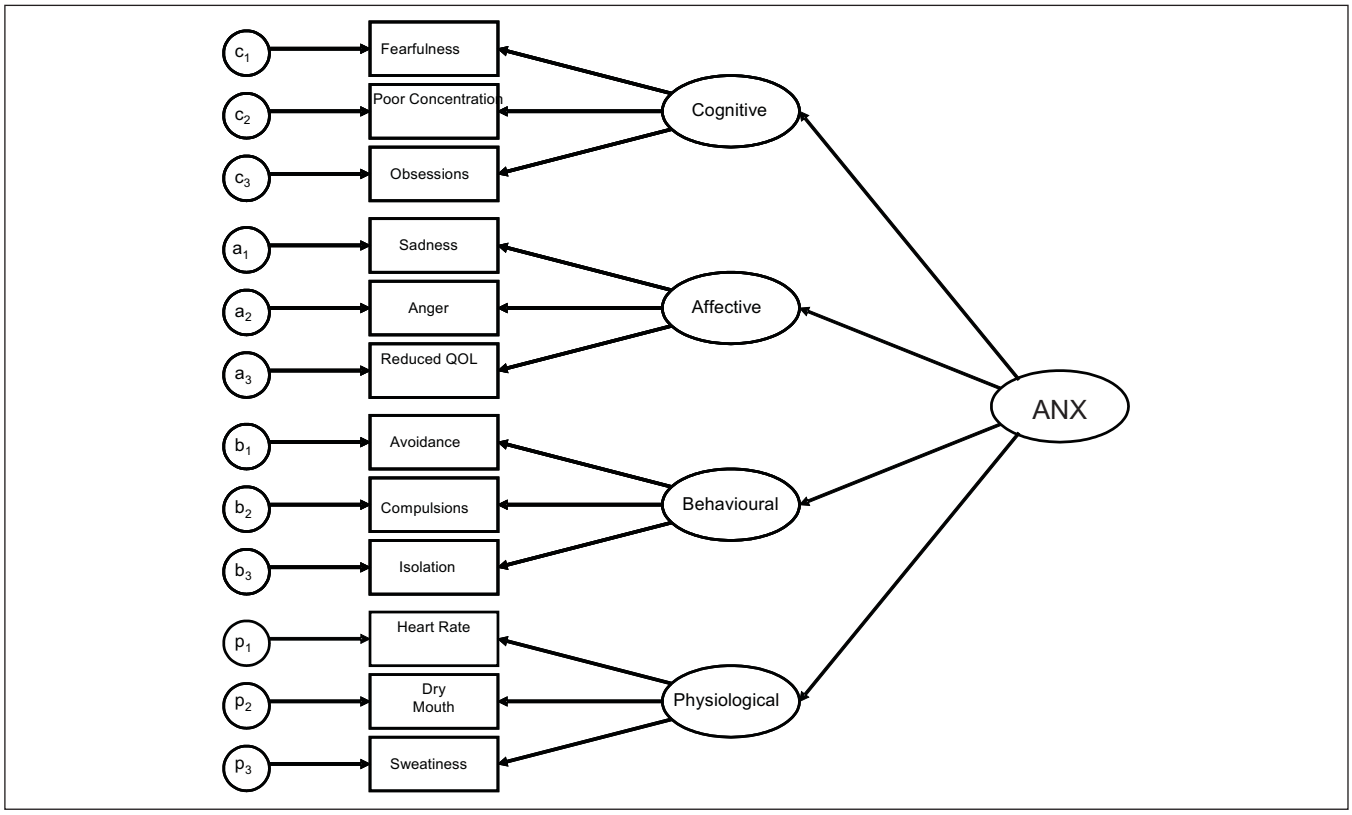
Confirmatory Factor Analysis

Until about 20 years ago, if someone said "factor analysis," there would be relatively little ambiguity about what he or she meant (3). It is a technique that is used when we have many items or variables and want to see whether they can be explained by a smaller number of factors. We enter the data, close our eyes, press the compute button, and see what comes out. That is, we don't have any a priori hypotheses regarding which variables or items will cluster together on the same factor. Even if we did have some hunches (for example, when we are analyzing a questionnaire we developed and have some idea of which items should tap the same construct), there is no way to tell this to the computer program ahead of time. All we can do is look at the output and say that the results are pretty close to what we expected, or we can go back to the drawing board and rewrite the items in the hope that the next iteration will give us results that are more to our liking. When CFA came upon the scene, there had to be some way to differentiate it from the more traditional form, so the older method was renamed EFA in recognition that we use it when we're simply examining the data.

As its name implies, CFA is used when we do have a priori hypotheses about which items or variables are grouped together as manifestations of an underlying construct and wish to test how well our data match—or fit—this model. As with path analysis, it is very helpful to draw the hypothesized relations in a diagram, particularly as the most commonly used computer programs, such as LISREL (SSI, Lincolnwood, IL), AMOS (SPSS, Chicago, IL), EQS (Multivariate Software, Encino, CA), and Mplus (Muthén & Muthén, Los Angeles, CA), accept these diagrams as input—it's not necessary to specify the relations mathematically. Let's, then, begin by drawing our theory of anxiety, which is shown in Figure 1.

If you recall the diagrams that were used with path analysis, you'll remember that there were 2 types of symbols: rectangles to represent the measured variables, both endogenous and exogenous; and circles to show the disturbance, or error, terms. In SEM (of which CFA is a subset), a third symbol is used: ovals, to depict the latent variables. Thus Figure 1 shows that there's a latent variable, anxiety, which in turn comprises 4 latent variables—cognitive, behavioural, affective, and physiological. These in turn give rise to several measured variables, each with an associated disturbance, or error, term. The figure shows that all 4 latent variables have 3 measured

Figure 1 A structural equation model of anxiety, with its 4 subcomponents and their measured variables



variables, but this was done simply because it was easier for me to draw it this way. In reality, each latent variable can have any number of measured variables, although, as I'll discuss later, there should ideally be at least 3.

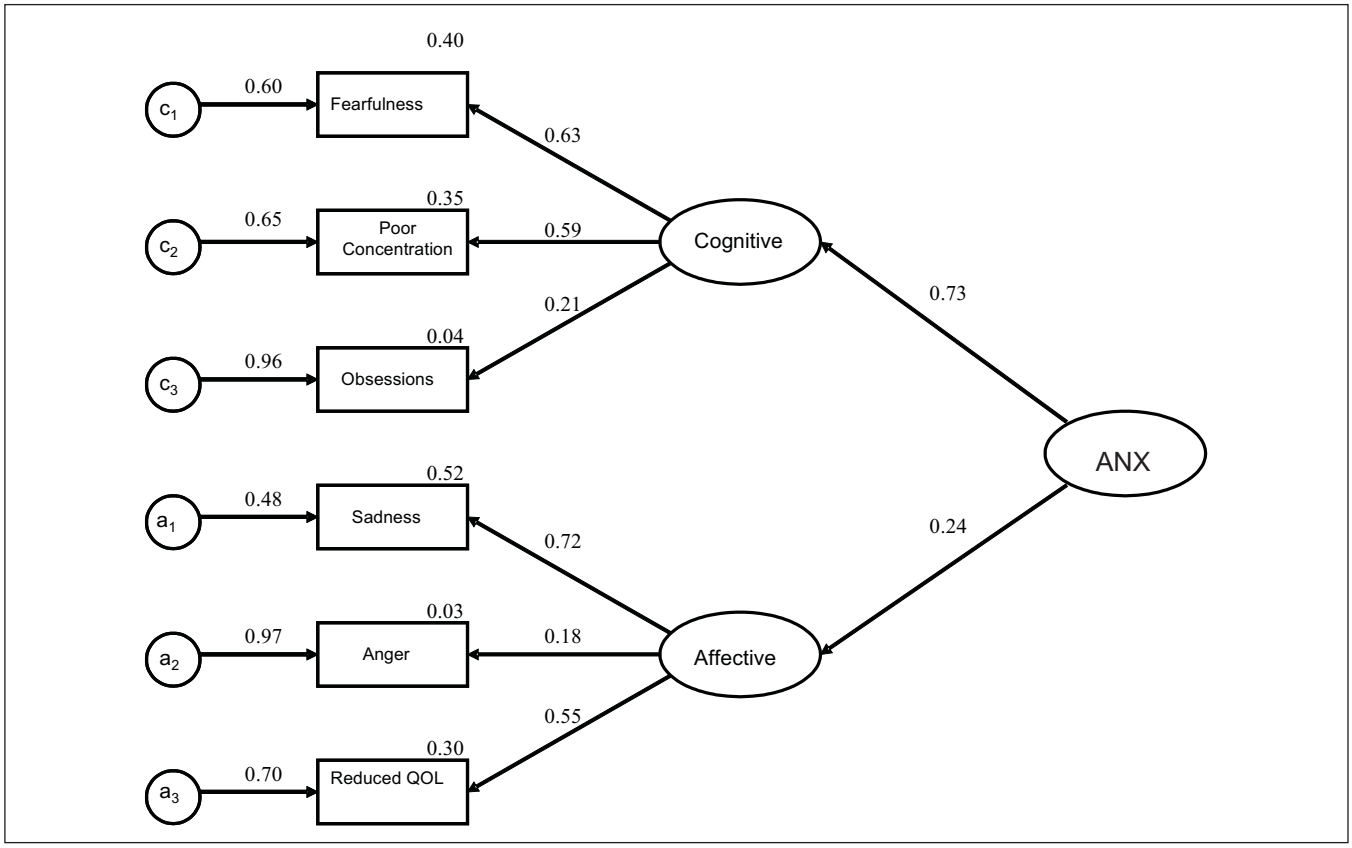
The direction of the arrows is important, not only for the analyses but also as a reflection of the underlying theory of latent variables, CFA, and SEM in general. The arrows from anxiety to the other latent variables, and from those 4 to the measured variables, mean that anxiety leads to 4 areas of involvement and that each of those gives rise to the observed variables. That is, if it weren't for the underlying construct of cognitive changes, for example, the observed behaviours would not be correlated with each other. Of course, they exist in people, but there would be no reason to expect that fearfulness and obsessions would go together, were it not for the fact that they are both outward manifestations of anxiety.

I mentioned in the paper on path analysis that it is a model testing procedure and should not be used for model building. The same injunction applies to SEM in general and CFA in particular. This is reflected in what CFA tells you, in contrast to what you are told with the exploratory form of factor analysis. In EFA, the program may say that variables A, C, and F, for example, belong together in factor Y, whereas variables B, D, and E load most highly on Factor Z. Even if you had hypothesized a different combination of variables clustering together, EFA will simply go with the math and show you the "best" configuration (where "best" may be defined differently in the various forms of EFA).

When you do a CFA, though, you stipulate where you think the variables should load, and the program tells you simply whether your model fits the data. If the model doesn't fit, there are few clues to guide you how to shuffle the variables around to make the model better fit the data. Further, even if the model does fit, that doesn't guarantee that some other way of arranging the variables (that is, a different model) would not lead to an even better fit. Thus your guide to the model is your theory, knowledge, or previous research, rather than reliance on statistical criteria.

CFA by itself is an extremely powerful and useful tool. For example, when validating a scale, EFA must rely on post hoc reasoning: "Yes, these results seem to make sense and more or less conform to what I had expected." With CFA, though, your hypotheses must precede the data analysis, so that a good

Figure 2 A subset of Figure 1, showing the path coefficients, squared multiple correlations, and error variances

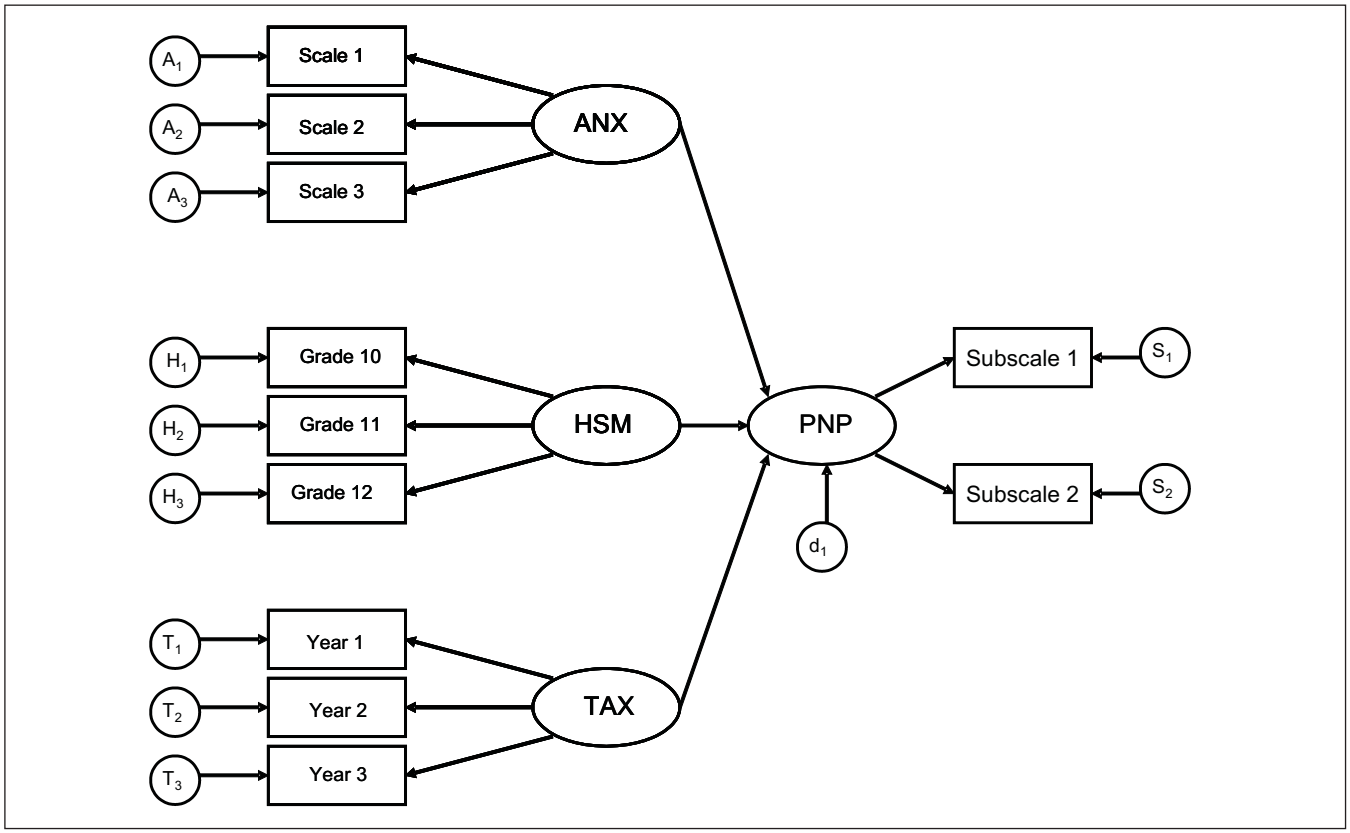


fit is even stronger evidence that the scale is structured as you thought.

Let's take a look at some (fictitious) results to see what CFA can tell us. After specifying the model, the computer will print a diagram similar to Figure 2. (I will not go into the messy details of how to specify the model because they vary from one program to the next and can get somewhat technical. With uncharacteristic modesty, let me recommend the chapter on SEM in *Biostatistics: The Bare Essentials* [4] as a place to start.) To simplify matters, Figure 2 shows only 2 of the 4 components of anxiety. Above each arrow from a latent variable to a variable is a number, called the path coefficient. This is equivalent to the factor loadings in EFA, so it can range from –1.0 to 1.0, with higher numbers (positive or negative) showing a stronger association. As can be seen, the variable "obsessions" doesn't fit very well with the cognitive trait; and the variable "anger" doesn't seem to go with the affective trait. The numbers over the variable names are the SMRs, which are simply the squared values of the path coefficients; they are interpreted in the same way as R^2 multiple regression—in terms of how much of the variance in one variable is explained by, or is in common with, the other variable. Finally, the numbers over the arrows between the error terms and the variables are the variances of the errors. You'll note that the sum of the SMR plus the error variance for each variable is 1.0; that is, all the variance of a variable is divided between that shared with the latent variable and error. This is equivalent in EFA to the communality (that portion of the variable's variance explained by the factors) and the uniqueness (what's left over); again, same concepts, different terms. At a higher level, note that, in this example, the cognitive domain is correlated more highly with the latent trait of anxiety than is the affective realm, as reflected in their respective path coefficients.

Another use of CFA is to compare the psychometric properties of different versions of a scale, or to determine whether it performs the same way with different groups. For example, to see whether men and women respond similarly to the items on a test, we can begin by doing an EFA on the women's data. The results of this EFA then constitute the model against which we test the data from the men. Then, we can run the CFA several ways, each time imposing stricter and stricter criteria for similarity. In our first run, we can simply see whether the same items load on the different factors. If this model fits, we can then add the restriction that the magnitudes of the

Figure 3 A structural equation model of the factors associated with PNP



factor loadings must be the same in both groups. If there is still a good fit, the final step would be to see whether the variances of the items are also similar across groups. Once a scale has passed these 3 increasingly rigourous tests, we can be fairly confident that it is performing in an equivalent manner across groups. The same approach is used to assess a translated form of a scale; it is compared against the factor structure in the original language. A nice recent example of this was done by Furukawa and others (5).

Structural Equation Modelling

With this background to CFA, and with the previous paper on path analysis (1), it is a relatively easy step to SEM. Instead of being limited to drawing paths among the measured variables, as we were with path analysis, we can draw paths among the latent variables. Each of the latent variables has at least 2 (and ideally 3 or more) associated measured variables, so that each latent variable becomes a small CFA in its own right. In fact, we'll use the same example of trying to predict a subject's degree of PNP on the basis of ANX, HSM, and TAX (1). Now, though, we'll treat each of these 3 as if they were latent variables, as shown in Figure 3. In keeping with the convention, what were squares in the original figure in the previous paper are now ovals, and each has a number of measured variables associated with it.

Here, ANX is measured by 3 different scales. Instead of using the marks from only 1 school year to measure HSM, we will use grades from 3 years and, similarly, look at errors for the past 3 years to measure TAX. PNP is a bit more difficult, because we have only one scale to measure it. For reasons that will be explained shortly, we randomly divide the scale into 2 parts, treating each as if it were a separate scale.

This step in SEM is called the model specification stage. Although no mathematics is involved, it is probably the most difficult—and most important—part. It is the most difficult because it requires the most thought and understanding of the theoretical model of the purported influences on PNP. No computer program can help us at this stage, only our knowledge of the field. It is the most important step because everything depends on how well we specify the model. The computer programs may help us in determining whether some variables aren't important, and as I explained in the previous article, we can play with different paths to see whether they improve the model. However, the primary cause of poorly fitting models (not only in SEM but also in path analysis and

Table 1 Correlations among 2 scales of anxiety (A_1 and A_2) and 2 scales of introversion (I_1 and I_2)

	A_2	I_1	I_2
A_1	0.74	0.49	0.42
A_2	1.00	0.45	0.40
I_1		1.00	0.70

multiple regression) is the omission of crucial variables, and there are no programs in the world that can help us in this regard. For example, if the prime determinant of PNP is actually the PNP level of one's parents (because of either genetics or learning), and if this isn't correlated with any of the other variables we're examining, our model will explain little of the variance, and we will never know why.

The next step is relatively easy: we simply run the computer program. Because our model is complex, so is the output. In essence, we are specifying 5 separate models: 4 CFAs (one for each of the latent variables) as well as the one that ties them all together. Before looking at the overall fit of the model, we should look at each of the CFAs. The main focus is on the paths—from the latent variable to the measured variables and from that latent variable to the next one in the path. Do they all have the right sign? Are they significant? If the answer to either question is no, it may be best to respecify the model by dropping nonsignificant variables and (or) seeing whether there are others in your data set that should be included.

Here, it is worth mentioning another advantage of CFA and why we prefer to deal with latent variables with 2—and ideally more—measured variables associated with them, rather than simply measured variables, as in path analysis. To keep the example simple, we'll deal with only 2 variables, anxiety and introversion, and—for reasons that will become obvious in a minute—measure each with 2 scales (A1 and A2 for anxiety; I1 and I2 for introversion). The usual way to test the hypothesis that the constructs are correlated with each other is to give the scales to a group of people and use Pearson's correlations. Table 1 presents the (again fictitious) results for 200 people.

As can be seen, the 2 anxiety scales are correlated 0.74 with each other, and the 2 introversion scales are correlated 0.70 with each other. The correlations between the anxiety and introversion scales range between 0.40 and 0.49, which is in the moderate range.

The problem, though, is that the magnitude of the correlations between the anxiety and introversion scales is affected by 3 factors: the degree to which these 2 constructs are actually related, the reliabilities of the anxiety scales, and the reliabilities of the introversion scales. Because the reliability of any scale is less than 1.0, the correlation that we find always underestimates the true correlation between the variables. We can get a better estimate of the true correlation if we disattenuate the variables, that is, if we compensate for the lack of perfect reliability (6)—but how do we know what the reliability is? By using several indices (or splitting each index in half, as we did with PNP), we can treat them as parallel forms of the same scale. In this example, even though they are different, each of the anxiety scales could be seen as (imperfect) measures of the trait of anxiety (and similarly for introversion). The degree to which the correlations between the 2 scales of a construct are less than 1 reflects the magnitude of this imperfection, that is, their parallel form reliability. In SEM, this is taken into account when the correlations among the latent variables ar examined; hence their "true" correlation is reflected.

When we rerun this problem, looking at the correlation between the latent variable of anxiety (measured with the 2 scales) and the latent trait of introversion (with its 2 scales), we find a correlation of 0.62, which is considerably stronger than the 0.40 to 0.48 we found previously and better reflects the actual relation between the traits.

Now let's return to the SEM example. Once we have cleaned up the model by pruning noncontributory paths, we can examine the fit of the overall model. The most common index of how well the data match the model (although not necessarily the best) is the χ^2_{GoF} (the chi-squared test for goodness-of-fit). Actually, the name is somewhat of a misnomer—it's really a badness-of-fit test. Usually, we want chi-squared to be statistically significant; in path analysis, though, we want χ^2_{GoF} to be nonsignificant. Why our change of heart? In general, chi-squared tests how much our data deviate from some hypothesized model. In the usual case that we're familiar with from introductory statistics, the model is that the variables are independent from one another, and we are delighted when we can reject this null hypothesis and conclude that the variables are in fact related. However, when we use the χ^2_{GoF} test in path analysis (or with other statistical tests), our hypothesized model is the one we have drawn (as opposed to the null hypothesis that nothing is related). If χ^2_{GoF} is statistically significant, that means the data differ from (that is, do not fit) the model, which is not what we want. Thus we want a path model that results in a nonsignificant χ^2_{GoF}. The good news is that all programs print out the results of this test and that the χ^2_{GoF}, in contrast to the tests I'll discuss next, has a probability level associated with it. The bad news is that we can't fully trust the results because they are highly dependent on the sample size. If the study has relatively few subjects, then χ^2_{GoF} may be nonsignificant even with a patently ridiculous model, simply because there isn't enough power to reject the null hypothesis.

Conversely, with a very large sample size, even minor and trivial deviations of the data from the model can result in statistical significance. Therefore, we should keep the results of the χ^2_{GoF} test in mind but not be overly influenced by it.

Another fit index is the RMSEA, which is a variant of the χ^2_{GoF} in that it sees how much the data deviate from the model. Values over 0.10 are considered to be a bad fit, those less than 0.08 reflect a reasonable fit, and values less than 0.05 indicate a good fit.

There are myriad other fit indices, all of which can be interpreted as measures of association or effect size (4). They can be grouped into 4 main categories (the fact that there are 4 categories indicates just how many individual indices there are and that none has been accepted as the gold standard). The comparative fit indices represent one type; these generally yield scores between 0 and 1. As the name implies, they show how good the model is, compared with some alternative. Most often, the alternative model is that all the variables are independent of one another, that is, that all the correlations (more accurately, the covariances) are zero. Because this is highly unlikely—let's not forget Meehl (7), who said that everything is correlated with everything else—it's not surprising that 0.90 is the minimally acceptable value, with 0.95 being the minimum if the χ^2_{GoF} test is significant (8). The second class of fit indices, which also have values between 0 and 1, reflect how much of the variance in the data can be accounted for by the model; again, 0.90 (or 0.95, if χ^2_{GoF} is significant) is the absolute minimum.

For both of these classes, there are modifications of the basic indices, reflecting their parsimony. These are based on the fact that 2 things happen as we add more variables. First, the amount of variance accounted for by the model increases, with rare exceptions. At the same time, each new variable also adds more error variance. Statistical techniques, though, cannot differentiate between true variance and error variance, so they find the best model that fits all of the variance. The problem this introduces is that, if we were to measure the exact same variables on a new sample of people, the true variance should be the same, but the pattern of the error variance would be quite different, since we assume that error is random. Consequently, the original model won't fit the new data as well. The parsimony indices penalize you for adding more variables, much as the adjusted R^2 in multiple regression imposes a penalty proportional to the number of variables in the model.

Unfortunately, we can only interpret the RMSEA and the other fit indices (with the exception of the χ^2_{GoF}) by using rules of thumb (under 0.08 for RMSEA, over 0.90 for the others if χ^2_{GoF} isn't significant, and over 0.95 if it is). There are no statistical tests of significance for these.

These techniques of path analysis, CFA, and SEM ask a lot from both the user and the reader. They introduce new terms for new concepts (for example, endogenous and exogenous variables and recursive and nonrecursive models), replace terms we know (construct or factor) with novel ones for the same concept (latent trait), and require specialized computer programs. What they give us in return, though, are more powerful ways of thinking about and analyzing our data—ways that more closely approximate the real world of many variables that interact in complex fashions that don't neatly fit into cause–effect relations.

Summary

In the previous article in this series (1), we saw how path analysis extended multiple regression by allowing chains of association between variables; for example, we saw how variables A, B, and C could affect variable D, which in turn influences variable E. A seemingly very different technique, EFA, was explained in another paper (3). CFA modifies EFA in that the user specifies a priori which items should load on which factors. Although this at first glance appears to be a limitation, demanding that the user have more information (or more sophisticated hunches) before he or she begins is in fact a major benefit, for 2 reasons. First, it yields better evidence that the composition of the scale matches one's assumptions, compared with having to rely on after-the-fact pleading that the results are sufficiently congruent. Second, it allows the user to compare the properties of the scale across populations or versions.

SEM both incorporates CFA and extends path analysis, by allowing the user to examine the relations among latent—that is, unseen but hypothesized—variables. Each latent variable has 2 or more measured variables associated with it. Thus each latent variable is a small CFA in its own right, testing the minihypothesis that the measured variables are in fact the measurable manifestations of the latent one. This also provides an added benefit in that the correlations among the measured variables are an indication of their reliability, and SEM can correct for this. Consequently, the relations among the latent variables reflect their true correlations uncontaminated by measurement error.

Funding and Support

No funding or other support was received for this paper.

Acknowledgement

I thank Dr Chittaranjan Andrade for his helpful comments on this and the previous paper in the series.

References

1. Streiner DL. Finding our way: an introduction to path analysis. Can J Psychiatry 2005;50:115–22.
2. Antony MM. Assessment of anxiety and the anxiety disorders: an overview. In: Antony MM, Orsillo SM, Roemer L, editors. Practitioner's guide to empirically based measures of anxiety. New York (NY): Kluver Academic/Plenum; 2001. p 7–17.
3. Streiner DL. Figuring out factors: the use and misuse of factor analysis. Can J Psychiatry 1994:39;135–40.
4. Norman GR, Streiner DL. Biostatistics: the bare essentials. 2nd ed. Toronto (ON): BC Decker; 2000.
5. Furukawa TA, Streiner DL, Azuma H, Higuchi T, Kamijima K, Kanba S, and others. Cross-cultural equivalence in depression assessment: Japan–Europe–North America study. Acta Psychiatr Scand 2005;112:279–85.
6. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford (UK): Oxford University Press; 2003.
7. Meehl P. Why summaries of research on psychological theories are often uninterpretable. Psychol Rep 1990;66:195–244.
8. Klein RB. Principles and practice of structural equation modeling. New York (NY): Guilford Publications; 1998.
9. Streiner DL. Figuring out factors: the use and misuse of factor analysis. Can J Psychiatry 1994;39:135–40.

Manuscript received July 2005, revised, and accepted October 2005.
This is the 25th article in the series on Research Methods in Psychiatry. For previous articles, please see Can J Psychiatry 1990;35:616–20, 1991;36:357–62, 1993;38:9–13, 1993;38:140–8, 1994;39:135–40, 1994;39:191–6, 1995;40:60–6, 1995;40:439–44, 1996;41:137–43, 1996;41:491–7, 1996;41:498–502, 1997;42:388–94, 1998;43:173–9, 1998;43:411–5, 1998;43:737–41, 1998;43:837–42, 1999;44:175–9, 2000;45:833–6, 2001;46:72–6, 2002;47:68–75, 2002;47:262–6, 2002;47:552–6, 2003;48:756–61, and 2005;50:115–122.
[1]Director, Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care, Toronto, Ontario; Professor, Department of Psychiatry, University of Toronto, Toronto, Ontario.
Address for correspondence: Dr DL Streiner, Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, ON M6A 2E1 dstreiner @ klaru-baycrest.on.ca

Résumé : Construire un meilleur modèle : une introduction à la modélisation des équations structurelles

L'analyse factorielle confirmatoire (AFC) et la modélisation des équations structurelles (MES) sont des extensions efficaces de l'analyse des pistes causales, qui a été décrite dans un précédent article de cette série. L'AFC diffère des analyses factorielles exploratoires plus traditionnelles en ce que les relations entre les variables sont spécifiées a priori, ce qui permet des tests plus puissants de la validité conceptuelle des échelles. Elle peut aussi servir à comparer différentes versions d'une échelle (par exemple, en anglais et en français), et à déterminer si l'échelle a un rendement équivalent dans différents groupes (par exemple, les hommes et les femmes). La MES développe l'analyse des pistes causales en permettant d'établir des pistes entre les variable latentes (lesquelles, dans d'autres techniques, se nomment concepts factoriels ou hypothétiques), c'est-à-dire les variables qu'on ne voit pas directement, mais plutôt par leur effet sur les variables observables, comme les questionnaires et les mesures du comportement. Chaque variable latente et ses variables mesurées associées forment de petites AFC, avec l'avantage ajouté que les corrélations parmi les variables peuvent être corrigées pour la non-fiabilité des mesures.