Advanced Quant Techniques

July 14, 2011

Structural Equation Modelling

Scott MacLean, Nulink Analytics

Event sponsored by Affinnova

For more information about Affinnova visit http://www. affinnova.com/

For more information about NewMR events visit newmr.org

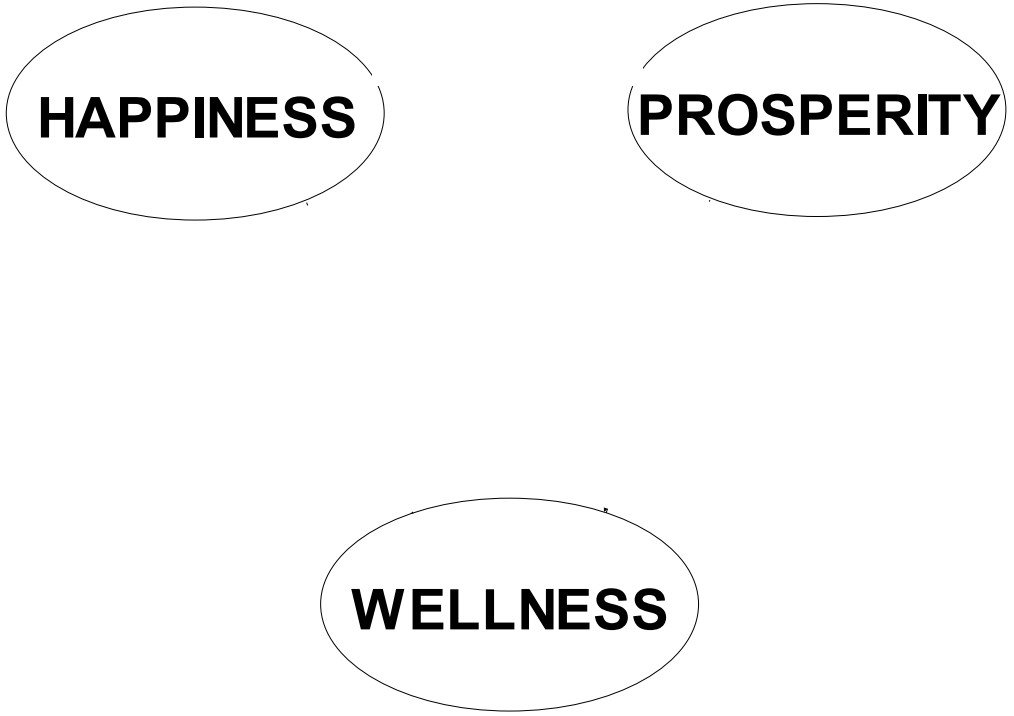
Scott MacLean

Nulink Analytics

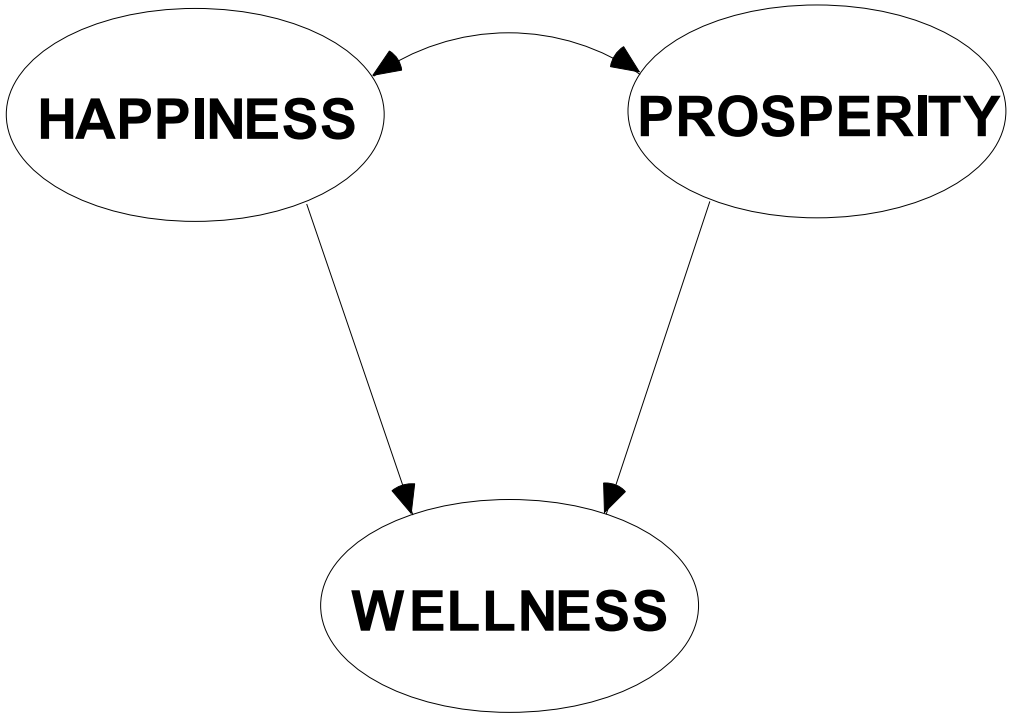
Structural Equation Modelling

- On the basis of things we can measure, we attempt to make predictions of things we cannot measure

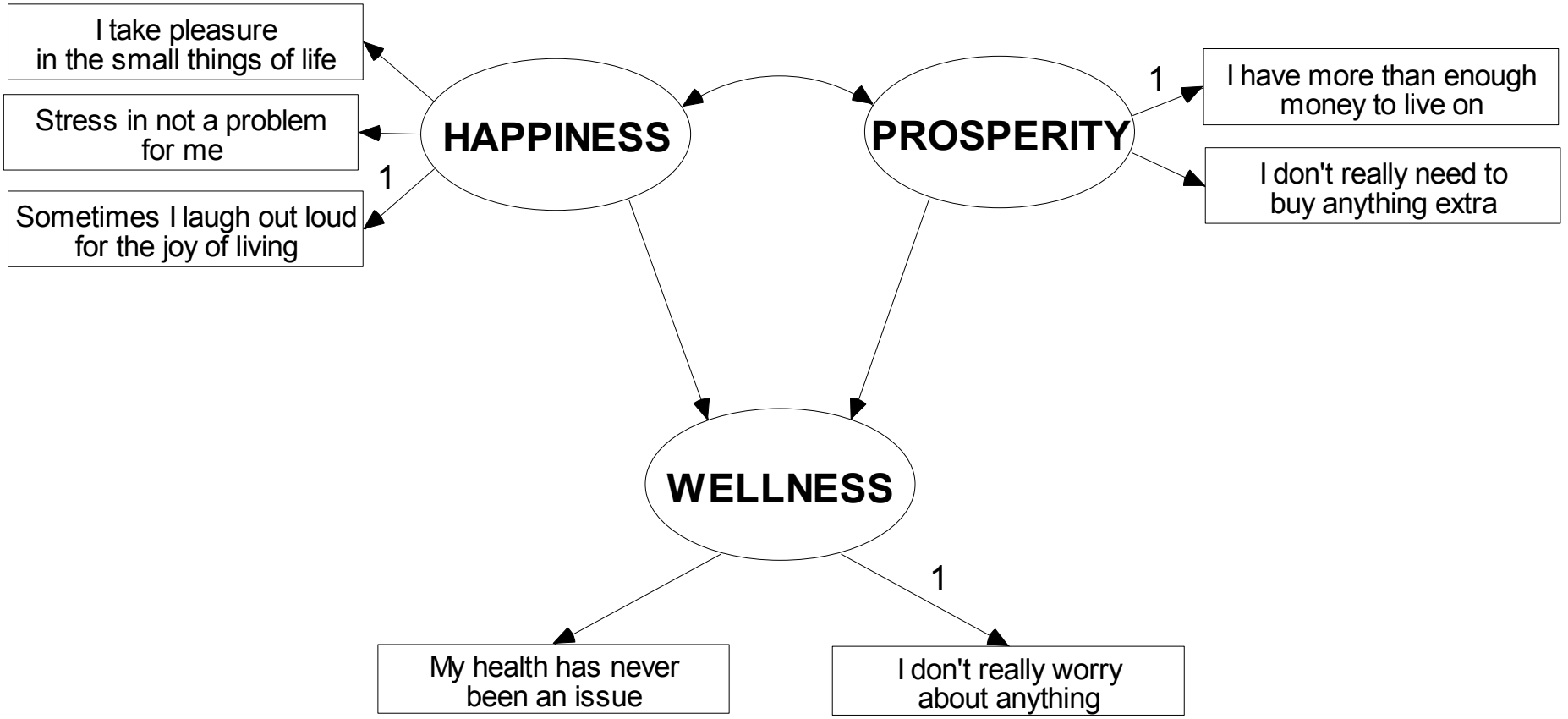
Start with some constructs/ factors



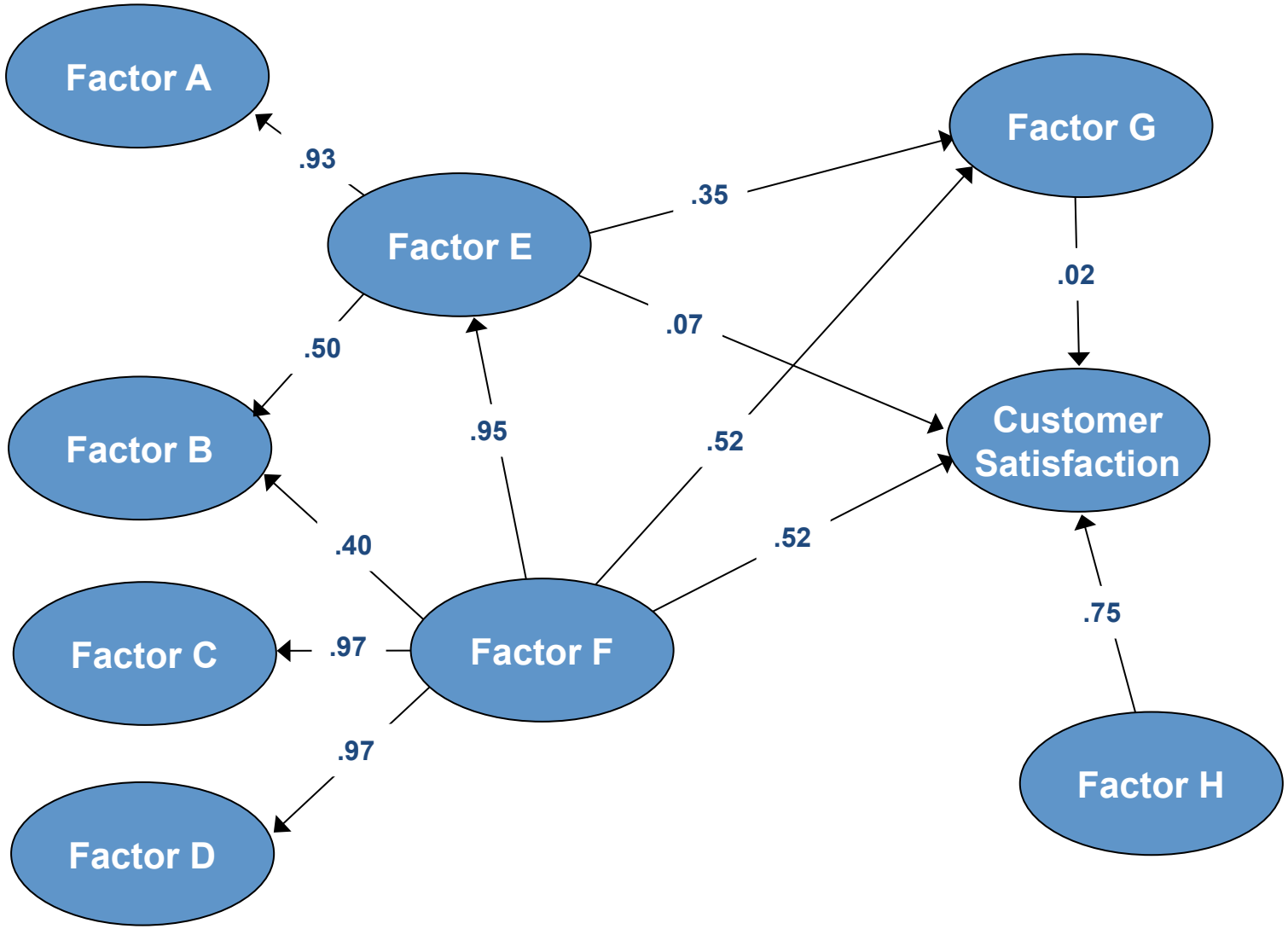
Add some links



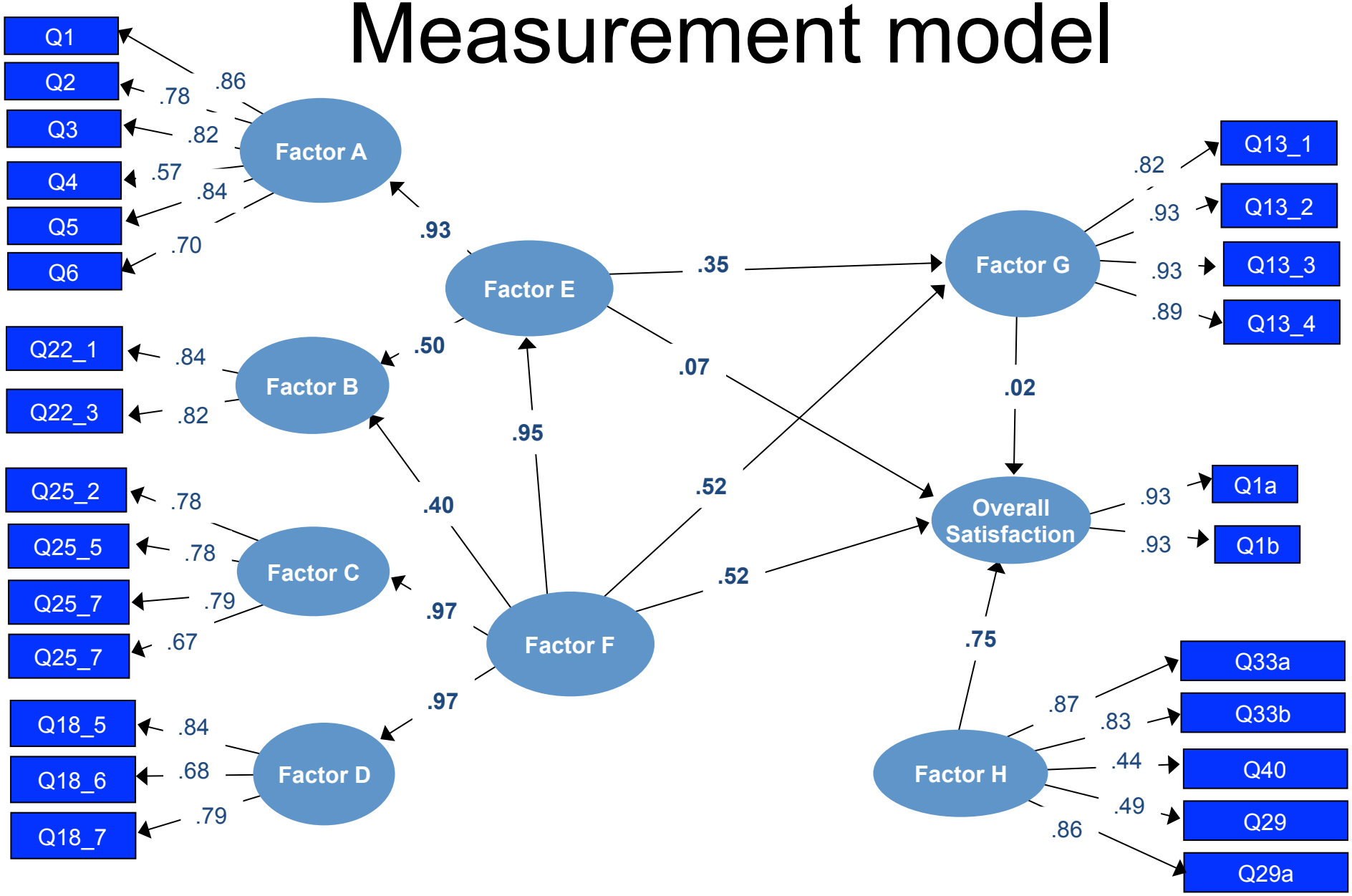
Add some indicator variables



Structural model example



Measurement model



Goodness of Fit

Goodness of fit indices abound in covariance-based SEM. No one fit index is best, and controversy over both which to use, and what values indicate lack of fit, is extensive.

Four of the most common are:

- **CFI** or Comparative Fit Index
- **GFI** or Goodness of Fit Index
- **RMSEA** or Root Mean Square Error of Approximation
- **cmin/df** or minimum chi square value divided by the degrees of freedom

These are discussed in the accompanying Glossary of Terms.

With PLS-based SEM, goodness of fit is not such an issue

Reflective versus Formative Indicators

- A construct should be modelled as having formative indicators if all of a number of conditions hold, eg.:
  - The indicators are viewed as defining characteristics of the construct
  - Changes in the indicators are expected to cause changes in the construct
  - Changes in the construct are not expected to cause changes in the indicators
  - The indicators do not necessarily share a common theme
  - Eliminating an indicator may alter the conceptual domain of the construct
  - A change in the value of one of the indicators is not necessarily expected to be associated with a change in all of the other indicators

Why is this important* ?

*"... **strong evidence** that measurement model misspecification of even one formatively measured construct within a typical structural equation model can have very **serious consequences** for the theoretical conclusions drawn from that model "*

***Jarvis, C. B., Mackenzie, Scott B. and Podsakoff, Philip M. (2003) *A critical review of construct indicators and measurement model misspecification in marketing and consumer research* Journal of Consumer Research, Vol 30 (September) 199-218.**

Not only that ...

“ The entire model could appear to adequately fit the data, even though the structural parameter estimates within that model exhibit very ***substantial biases*** *that would result in* ***erroneous inferences*** *...”*

... but also ...

" … paths emanating from a construct with a misspecified measurement model are likely to be ***substantially inflated*** *...*

… paths leading into a construct with a misspecified measurement model are likely to be ***deflated*** *"*

Is it a big issue ?

- Four journals were searched* for the **24-year** period from 1977 through 2000
- Aim – to identify all empirical applications of latent variable SEM or confirmatory factor analysis
  - *Journal of Consumer Research*
  - *Journal of Marketing*
  - *Journal of Marketing Research*
  - *Marketing Science*

***Jarvis, C. B., Mackenzie, Scott B. and Podsakoff, Philip M. (2003) *A critical review of construct indicators and measurement model misspecification in marketing and consumer research* Journal of Consumer Research, Vol 30 (September) 199-218.**

What did they find?

- 178 articles containing 1,192 constructs modelled as latent factors with multiple indicators

	Should have been modelled as reflective	Should have been modelled as formative	TOTAL
Was modelled as reflective	810	336	1,146
Was modelled as formative	17	29	46
TOTAL	827	365	1,192

And it means ...

Over a 24 year period in academic, refereed papers appearing in four of the world's leading market research / marketing science journals, **nearly 30% of modelled constructs were wrong**

Formative example

- Customer complaints, where the indicators might be:
  - Frequency of complaining to store manager
  - Incidence of telling friends and relatives about a bad service experience
  - Likelihood of reporting the supplier to a consumer complaints agency
  - Likelihood of pursuing legal action against the supplier
  - etc.
- A high score on one item would certainly influence the level of the latent construct, but would not necessarily have an effect on the other items

Analysis: formative vs reflective

- It can be difficult to analyse models that contain formative factors, unless the PLS-based approach to SEM is utilised
- The covariance-based can be used, but only under very specific circumstances*

* See http://nulinkanalytics.com.au/D2-S2-TM1-03%20Scott%20MacLean.pdf for further details

Software suggestions

- AMOS
  - Available from http://www-01.ibm.com/software/analytics/spss/products/statistics/amos/
- SmartPLS
  - Beta release available from www.smartpls.de

Further reading suggestions

- http://nulinkanalytics.com.au/D2-S2-TM1-03%20Scott%20MacLean.pdf
- http://www.smallwaters.com/whitepapers/marketing/
- "Structural Equation Modelling in Marketing – Parts I, II and III" in Dictionary of Quantitative Management Research; **Sage Publications Ltd**; ISBN: 9781412935296

Training suggestions

- A 3-day Course on PLS Path Modelling and XLSTAT-PLSPM software: Basic Concepts and Foundations, Advances and Applications by W.W. Chin, V. Esposito Vinzi, M. Tenenhaus 26-27-28 September 2011, Paris, France http://www.xlstat.com/en/training/ & http://www.xlstat.com/file/PLSPMCourse.pdf

- Online R and PLS courses for Aug 2011 through June 2012, https://www.regonline.com/2-for-1

Thank you

Scott MacLean

Nulink Analytics

Q & A

Scott MacLean
Nulink Analytics

Sue York
The Future Place

Scott MacLean

Scott MacLean
Director, Nulink Analytics
T: +61 419 504 588
E: scott@nulinkanalytics.com
W: www.nulinkanalytics.com

Glossary of terms*

Indicators are observed variables, sometimes called manifest variables or reference variables, such as items in a survey instrument.

Latent variables are the unobserved variables or **constructs** or **factors** which are measured by their respective indicators.

The **measurement model** is that part (possibly all) of a SEM model which deals with the latent variables and their indicators.

Exogenous variables are independents with no prior causal variable (though they may be correlated with other exogenous variables, depicted by a double-headed arrow -- note two latent variables can be connected by a double-headed arrow (correlation) or a single-headed arrow (causation) but not both.

Endogenous variables are mediating variables (variables which are both effects of other exogenous or mediating variables, and are causes of other mediating and dependent variables), and pure dependent variables.

* Source: Garson, G. David (2007). "Structural Equation Modeling", in *Statnotes: Topics in Multivariate Analysis,* retrieved 21 August 2007 from http://www2.chass.ncsu.edu/garson/pa765/sructur.htm

Glossary of terms*

Path coefficients are the effect sizes calculated by the model estimation program. Often these values are displayed above their respective arrows on the arrow diagram specifying a model. In AMOS, these are labeled "regression weights," which is what they are, except that in the structural equation there will be no intercept term (unlike conventional regression, where this is normally specifically provided for).

Standardised (path) coefficients measure the extent to which a change of one standard deviation in the independent or predictor variable is reflected by a change in the dependent variable, also measured in units of its standard deviation.

Residual terms essentially measure the difference between the predicted and estimated actual values for the latent variables

Error terms essentially measure the difference between the observed and predicted values for the indicator variables

* Source: Garson, G. David (2007). "Structural Equation Modeling", in *Statnotes: Topics in Multivariate Analysis,* retrieved 21 August 2007 from http://www2.chass.ncsu.edu/garson/pa765/sructur.htm

Glossary of terms* (cont.)

Goodness of fit indices abound in Structural Equation Modelling. No one fit index is best, and controversy over both which to use, and what values indicate lack of fit, is extensive.

Four of the most common are:

- **CFI** or Comparative Fit Index
- **GFI** or Goodness of Fit Index
- **RMSEA** or Root Mean Square Error of Approximation
- **cmin/df** or minimum chi square value divided by the degrees of freedom

These are briefly discussed on the following pages.

Degrees of freedom (df) is a measure of the number of independent pieces of information on which the precision of a parameter estimate is based. The degrees of freedom for an estimate equals the number of observations (values) minus the number of additional parameters estimated for that calculation. As we have to estimate more parameters, the degrees of freedom available decreases.

* Source: Garson, G. David (2007). "Structural Equation Modeling", in *Statnotes: Topics in Multivariate Analysis,* retrieved 21 August 2007 from http://www2.chass.ncsu.edu/garson/pa765/sructur.htm

Glossary of terms* (cont.)

CFI compares the existing model fit with a null model which assumes the latent variables in the model are uncorrelated (the "independence model").

That is, it compares the covariance matrix predicted by the model to the observed covariance matrix, and compares the null model (covariance matrix of 0's) with the observed covariance matrix, to gauge the percent of lack of fit which is accounted for by going from the null model to the researcher's SEM model.

CFI is among the measures least affected by sample size.

CFI varies from 0 to 1 (if outside this range it is reset to 0 or 1). CFI close to 1 indicates a very good fit.

By convention, CFI should be equal to or greater than (about) .90 to accept the model, indicating that 90% of the covariation in the data can be reproduced by the given model.

* Source: Garson, G. David (2007). "Structural Equation Modeling", in *Statnotes: Topics in Multivariate Analysis,* retrieved 21 August 2007 from http://www2.chass.ncsu.edu/garson/pa765/sructur.htm

Glossary of terms* (cont.)

GFI varies from 0 to 1 but theoretically can yield meaningless negative values.

A large sample size pushes GFI up. Though analogies are made to R-square, GFI cannot be interpreted as percent of error explained by the model. Rather it is the percent of observed covariances explained by the covariances implied by the model.

That is, R-square in multiple regression deals with error variance whereas GFI deals with error in reproducing the variance-covariance matrix.

As GFI often runs high compared to other fit measures, some suggest using .95 as the cutoff.

By convention, GFI should by equal to or greater than (about) .90 to accept the model.

* Source: Garson, G. David (2007). "Structural Equation Modeling", in *Statnotes: Topics in Multivariate Analysis,* retrieved 21 August 2007 from http://www2.chass.ncsu.edu/garson/pa765/sructur.htm

Glossary of terms* (cont.)

RMSEA - by convention, there is good model fit if RMSEA less than or equal to .05. There is adequate fit if RMSEA is less than or equal to (about) .08.

It is one of the fit indexes less affected by sample size, though for smallest sample sizes it overestimates goodness of fit.

It may be said that RMSEA corrects for model complexity, as shown by the fact that df is in its denominator. However, degrees of freedom is an imperfect measure of model complexity. Since RMSEA computes average lack of fit per degree of freedom, one could have near-zero lack of fit in both a complex and in a simple model and RMSEA would compute to be near zero in both, yet most methodologists would judge the simpler model to be better (in other words, there is no compelling reason to accept a more complex explanation when a simpler explanation achieves the same outcome).

* Source: Garson, G. David (2007). "Structural Equation Modeling", in *Statnotes: Topics in Multivariate Analysis,* retrieved 21 August 2007 from http://www2.chass.ncsu.edu/garson/pa765/sructur.htm

Glossary of terms* (cont.)

cmin - the model chi-square, also called *discrepancy* or the *discrepancy function*, is the most common fit test, printed by all computer programs. The chi-square value should not be significant if there is a good model fit, while a significant chi-square indicates lack of satisfactory model fit, ie. the given model's covariance structure is significantly different from the observed covariance matrix.

There are three ways, listed below, in which the chi-square test may be misleading. Because of these reasons, many researchers who use SEM believe that with a reasonable sample size (eg. n > 200) and good approximate fit as indicated by other fit tests (eg. CFI, RMSEA), the significance of the chi-square test may be discounted and that a significant chi-square is not a reason by itself to modify the model.

- The more complex the model, the more likely a good fit.
- The larger the sample size, the more likely the rejection of the model
- The chi-square fit index is also very sensitive to violations of the assumption of multivariate normality.

* Source: Garson, G. David (2007). "Structural Equation Modeling", in *Statnotes: Topics in Multivariate Analysis,* retrieved 21 August 2007 from http://www2.chass.ncsu.edu/garson/pa765/sructur.htm

Glossary of terms* (cont.)

cmin/df – is the chi-square fit index divided by degrees of freedom, in an attempt to make it less dependent on sample size.

Some authors state that relative chi-square should be in the 2:1 or 3:1 range for an acceptable model.

Still others would argue that 3 or less is acceptable, and some allow values as high as 5 to consider a model adequate fit, while others insist relative chi-square be 2 or less.

Because of the inadequacies of cmin as a measure of goodness of fit (see previous page) we would mostly rely on the other fit measures for assessing the appropriateness (or, more accurately, lack of non-appropriateness) of the model.

* Source: Garson, G. David (2007). "Structural Equation Modeling", in *Statnotes: Topics in Multivariate Analysis,* retrieved 21 August 2007 from http://www2.chass.ncsu.edu/garson/pa765/sructur.htm

Glossary of terms* (cont.)

Collinearity – Collinearity (or multicollinearity – the two terms are interchangeable) occurs because two (or more) variables are related – they essentially measure the same thing.

Why is multicollinearity a problem? If the goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem. The predictions will still be accurate, and the overall r-square quantifies how well the model predicts the Y values.

However, if the goal is to understand how the various X variables impact Y, then multicollinearity is a big problem. One problem is that the regression coefficients can be indicated as statistically insignificant, even though the variables concerned are important. The second problem is that the confidence intervals on the regression coefficients will be very wide. The confidence intervals may even include zero, which means we can't even be confident whether an increase in the X value is associated with an increase, or a decrease, in Y. Because the confidence intervals are so wide, excluding a subject (or adding a new one) can change the coefficients dramatically – and may even change their signs.

* Source: http://www.graphpad.com/articles/Multicollinearity.htm

Glossary of terms* (cont.)

(Exploratory) Factor Analysis – Factor Analysis is used to uncover the latent structure (dimensions) of a set of variables. It reduces attribute space from a larger number of variables to a smaller number of factors and as such is a "non-dependent" procedure (that is, it does not assume a dependent variable is specified).

Factor Analysis can be used for any of a number of purposes, eg:

- To reduce a large number of variables to a smaller number of factors for modeling purposes, where the large number of variables precludes modeling all the measures individually
- To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component factors
- To create a set of factors to be treated as uncorrelated variables as one approach to handling multicollinearity in such procedures as multiple regression

* Source: Garson, G. David (2007). "Factor Analysis", in *Statnotes: Topics in Multivariate Analysis,* retrieved 18 September 2007 from http://www2.chass.ncsu.edu/garson/pa765/factor.htm