WEB PAGE FOR CHAPTER 16

MULTIPLE CHOICE QUESTIONS – SET A

1 With a scattergraph diagram, the dependent variable is represented on which axis?
(a) X
(b) Y
(c) Z
(d) P
(e) R

2 With reference to the regression equation, b_0 represents the:
(a) lack of gradient
(b) point at which the line of best fit intersects the Y axis
(c) slope of the line
(d) dependent variable
(e) independent variable

3 In a study relating the role of interest rates in determining the amounts invested by the public in bank term deposits:
(a) the amount invested is the dependent variable
(b) in a scattergraph, the interest rate would be placed on the Y axis
(c) the independent variable is represented on the Y axis
(d) in a scattergraph, the amounts invested are represented on the X axis
(e) none of the above

4 If the regression equation is: $Y = 1 + 2X$ where Y represents sales in thousands of dollars and X represents the number of customers:
(i) What would be the predicted level of sales if there were 20 customers?
(a) 41
(b) 60
(c) $41 000
(d) $60 000
(e) $42 000

(ii) How many customers would be predicted to be needed if a company wanted sales of $61, 000?
(a) 20
(b) 21
(c) 31
(d) 61
(e) 30

5 To obtain the Co-efficient of Multiple Determination, it is necessary to:
(a) take the square root of ‘r’
(b) compute the exponential value of ‘r’
(c) square ‘R’
(d) $r \times r \times r$
(e) take the square root of ‘R’

6 If the Co-efficient of Determination is equal to 0.81 this indicates that:
(a) 1.8% of the variation in X and Y is common
(b) 8.1% of the variation in X and Y is common
(c) .81% of the variation in X and Y is common
(d) 9% of the variation in X and Y is common
(e) 81% of the variation in X and Y is common

7 With reference to the regression equation, b_1 represents the:
(a) the independent variable
(b) point at which the line of best fit intersects the Y axis
(c) slope of the line
(d) dependent variable
(e) none of these

8 Identify the slope in the following equation: $Y = -3X + 7$:
(a) 7
(b) X
(c) –3
(d) –3X

9 Given a linear equation $Y = 2X - 7$ determine the value of Y when $X = 10$
(a) 13
(b) 3
(c) 10
(d) 14

10 Given a regression equation $Y = -5 + 2X$, which of the following is true?
(a) the correlation is negative
(b) the line slopes from top left to bottom right
(c) Y values are negative
(d) the intercept cuts the Y axis below the X axis

11 The extent to which observed values differ from their predicted values on the regression line is measured by the:
(a) intercept
(b) beta coefficient
(c) standard error of estimate
(d) line of least squares

12 Identify the independent variable where a person's age is used to predict life expectancy and therefore the cost of their life insurance:
(a) the person
(b) age
(c) cost
(d) life expectancy

13 The line of best fit:
(a) maximizes the correlation between X and Y
(b) minimizes the correlation between X and Y
(c) minimizes the distances between scatterpoints and the regression line
(d) minimizes the number of points the line misses.

SPSS ACTIVITIES

1 Access SPSS Chapter 16 Data File A and perform a linear regression using number of employees in each branch as a predictor of mean monthly sales values.

2 Access SPSS Chapter 16 Data File A, choose any other predictor (IV) and perform a linear regression to determine its ability to predict mean monthly sales value.

3 Access SPSS Chapter 16 Data File A. Conduct a multiple regression. We want to see whether mean monthly sales value is predicted by a combination of floor space and number of employees. Express the regression equation. Is this prediction higher than that from just number of employees alone?

4 Access SPSS Chapter 16 Data File A. Conduct a multiple regression entering all the predictors and determine the regression equation. Which are the best predictors and which contribute very little to the estimate of mean monthly sales value? Repeat with a heirarchical regression and a stepwise regression, and check whether the results appear consistent across the different methods.

5 Access SPSS Chapter 16 Data File B. Conduct a hierarchical and a stepwise multiple regression. Determine the best set of predictors for the DV profit in 2007. Interpret your results.

MULTIPLE CHOICE QUESTIONS – SET B

1 The meaning of the Co-efficient of Determination is that variation in the dependent variable can be explained by:
(a) the proportion of Y
(b) the variation in the independent variable Y
(c) the variation in the independent variable X
(d) all of the above
(e) none of the above

2 The Co-efficient of Determination is always ………………. and is often reported as a ………………….
(a) positive, portion
(b) negative, percentage
(c) positive, negative
(d) negative, positive
(e) positive, percentage

3 If the slope in a linear equation is positive, then the graph of the equation will be a line sloping from lower left to upper right:
(a) false
(b) depends on the intercept
(c) depends on the correlation
(d) true

4 If the dependent variable increases as the independent variable increases in an estimating equation, the correlation coefficient will be in the range:
(a) 0 to – 1
(b) 0 to + 1
(c) 0 to + .5
(d) 0 to – .5

5 If 64% of the variation in the dependent variable is explained by the regression line then the value of r is:
(a) +.8
(b) +.64
(c) +.64
(d) +.08

6 The regression equation for the relationship between age and total questionnaire score on attitudes to nuclear power (with high scores positive attitude) is:
nuclear power attitudes = 6.9643 + .6230 age, and r = +.6
(a) Explain what 6.9643 means
(b) Explain what .6230 means
(c) What is the likely score of someone aged 18?
(d) What is the likely score of someone aged 45?
(e) How well is variable Y explained by variation in variable X?

7 In linear regression the residuals are:
(a) the correlation between the predicted and actual scores
(b) the actual scores minus the predicted scores
(c) the actual scores plus the predicted scores
(d) the difference between the actual score and the predicted score

The following questions 8 – 9 relate to data in the tables below. The multiple regression concerns the prediction students' heights from those of both parents.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	dadheight, momheigh t[a]		Enter

[a] All requested variables entered.
[b] Dependent Variable: HEIGHT

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.458[a]	.210	.200	3.5119	.210	21.628	2	163	.000

[a] Predictors: (Constant), dadheight, momheight

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	533.498	2	266.749	21.628	.000[a]
	Residual	2010.335	163	12.333		
	Total	2543.833	165			

[a] Predictors: (Constant), dadheight, momheight
[b] Dependent Variable: HEIGHT

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	23.322	6.778		3.441	.001
	momheight	.301	.093	.234	3.246	.001
	dadheight	.354	.076	.338	4.686	.000

a. Dependent Variable: HEIGHT

8 The adjusted R^2 value tells us that:
(a) 40% of the variance in offspring height is explained by the combination of parental heights
(b) 20% of the variance in offspring height is explained by the combination of parental heights
(c) 21% of the variance in offspring height is explained by the combination of parental heights
(d) 45.8% of the variance in offspring height is explained by the combination of parental heights

9 The addition of the second IV adds:
(a) a non-significant F change
(b) a significant F change
(c) a significant R change
(d) a significant R^2 change

10 The F in the ANOVA table tells us:
(a) the DV is significant
(b) the IV's are significantly correlated
(c) the multiple correlation is significant
(d) the two IV's used together are significantly related to the DV

11 The coefficient table tells us:
(a) that the two IV's are individually both significant contributors to the prediction of the DV
(b) that momheight is a more significantly predictor than dadheight
(c) that both IV's contribute equally to the predictive power of the regression equation
(d) that the constant is the most important predictor as it has the biggest b value

12 The regression formula for prediction of student offspring height among this sample is:
(a) Student offspring height = 23.322 + .301 + .354
(b) Student offspring height = 23.322 + (.301)(.354)
(c) Student offspring height = 23.322 + .301momheight + .354dadheight
(d) Student offspring height = 23.322 (.301momheight + .354dadheight)

ADDITIONAL QUESTIONS

1 Given the following regression equation, estimate Y when X = 50.
$Y = 1.45 + .5(X)$

2 Given a correlation of +.8, a SD of 3 for Y and a score of 10 on X with a regression equation of $Y = 1.6 + 2.5X$:
(a) calculate the standard error of the estimate for Y and the 95% confidence interval for the predicted Y score
(b) If the correlation is now +.3 recalculate the standard error of the estimate and the 95% confidence interval
(c) Explain the difference between (a) and (b)

3 Consider the following regression equation:

$Y = 7.3 + 2.3X_1 + 4.1X_2 - 1.4X_3$ $R^2 = .78$ $F = 21.43$ $p<.01$

(a) What value does Y exhibit if $X_1 = 9$, $X_2 = 22$, and $X_3 = 17$

(b) How confident are you about this prediction?

(c) What does the negative sign for X_3 mean?

4 Consider the following regression equation:

$Y = 9.5 + 2.6X_1 + 4.8X_2 - 1.1X_3$

(a) What value would Y have if $X_1 = 7$, $X_2 = 20$, and $X_3 = 10$?

(b) Which of the three independent variables exhibits the largest effect on Y?

5 Two companies supply laboratory animals for the biology department. Company A sells rats at $6 each and charges a $10 delivery fee. Company B sells rats at $5 each but has a $20 delivery charge. In each the delivery charge is a one-off fee and does not relate to the number of rats delivered.

(a) for each company what is the linear equation that defines the total cost (Y) as a function of the number of rats (X)?

(b) If you are buying 20 rats which company provides the better deal?

6 The Emergency Services manager is concerned about response times to fires. He orders an investigation to determine if distance to the fire measured in km can explain response time measured in minutes. Based on 37 emergency calls, the following regression equation was compiled. What is the estimate of response time for calls 8 km and 20 km from the station?

$Y = 4.05 + 1.23(X)$

CLASS ACTIVITIES AND DISCUSSIONS

1 Explain the difference between correlation and regression to one of your classmates.

2 Either individually or as a group, collect data on variables you feel are related in the lifestyle of your fellow students (e.g. *number of hours spent studying each week, number of hours per week spent watching TV, number of hours per week surfing the internet, amount spent on mobile calls per week, distance lived from college amount spent weekly on transport to college, time spent weekly on journey to college, etc*.). For Y choose some variable like weekly cost of living (how much spent) or intention to leave university before completion of degree on 5 point scale.

Collect this data for at least 25 students. Compute the multiple regression equation. Are any predictor variables of no importance? Estimate some value of Y given some X values.

3 Explain to another class member under what conditions hierarchical and stepwise regression are used.

4 In groups, inspect the following tables then answer the following questions. Share your findings:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Previous Experience (months)[a]	.	Enter

[a] All requested variables entered.

[b] Dependent Variable: Beginning Salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.045[a]	.002	.000	$7,870.942

[a] Predictors: (Constant), Previous Experience (months)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	59692532	1	59692531.85	.964	.327[a]
	Residual	2.92E+10	472	61951721.26		
	Total	2.93E+10	473			

[a] Predictors: (Constant), Previous Experience (months)
[b] Dependent Variable: Beginning Salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	16690.478	490.646		34.017	.000
	Previous Experience (months)	3.397	3.460	.045	.982	.327

[a] Dependent Variable: Beginning Salary

(a) Name the IV and DV. Devise a null hypothesis.
(b) On the basis of the evidence is the null hypothesis rejected or retained? Justify your answer with evidence from three tables.

5 In groups, inspect the following tables and answer the questions below. Share your findings:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Previous Experience (months)[a]		Enter

[a] All requested variables entered.
[b] Dependent Variable: Current Salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.097[a]	.009	.007	$17,012.353

[a] Predictors: (Constant), Previous Experience (months)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.31E+09	1	1310179340	4.527	.034[a]
	Residual	1.37E+11	472	289420161.2		
	Total	1.38E+11	473			

a. Predictors: (Constant), Previous Experience (months)
b. Dependent Variable: Current Salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	35945.029	1060.488		33.895	.000
	Previous Experience (months)	-15.913	7.479	-.097	-2.128	.034

a. Dependent Variable: Current Salary

(a) Name the IV and DV?
(b) Express a null hypothesis.
(c) On the basis of the evidence is the null hypothesis retained or rejected?
(d) Formulate the regression equation
(e) What is the implication of the negative B?

ANSWERS TO REVIEW QUESTIONS

Qu. 16.1 (b)

Qu. 16.2 Y = –5 ; –1 ; +3 ; +13

Qu. 16.3 (c)

Qu. 16.4 a = \$30; b = \$630

Qu. 16.5 40 ± 2.94

Qu. 16.6 Explained = 87% approx.; unexplained = 13% approximately

Qu. 16.7 r = .79

Qu. 16.8 (c)

Qu. 16.9 (a) Y = 31.78 + 1.67(X); (b) = 0.918 = 91.8% variance in Y explained by variance in X; predicted score = 65, and this varies from the real score since ‘r’ and ‘r^2’ are not perfect.

ANSWERS TO ADDITIONAL QUESTIONS

1 26.45

2 (a) 10 ± 3.53; (b) 10 ± 5.58; (c) lower r means poorer prediction therefore greater range

3 (a) 94.4; (b) R^2 suggests that 78% of variation in Y can be predicted form variation in IVs combined and F is also significant so prediction is reasonable. (c) for every unit change in X, Y will decrease by 1.4

4 (a) 112.7; (b) X_2

5 (a) Y = 10 + 6X for company A; Y = 20 + 5X for company B
(b) For company A, Y = $130 and for company B, Y = $120. Company B provides best value

6 13.89 minutes; 28.65 minutes

ANSWERS TO MULTIPLE CHOICE QUESTIONS – SET A

1 (b), 2 (c), 3 (a), 4 (I c; ii e), 5 (c), 6 (e), 7 (c), 8 (a), 9 (a), 10 (d), 11 (c), ` 12 (b), 13 (c)

ANSWERS TO MULTIPLE CHOICE QUESTIONS – SET B

1 (c), 2 (e), 3 (d), 4 (b), 5 (a), 6 (a = constant; b = slope; c = 18; d = 35; e = 36% quite weak), 7 (d), 8 (b), 9 (b), 10 (d), 11 (a), 12 (c)

ADDITIONAL MATERIAL

Path analysis and causal modelling: Advanced applications of correlation and regression

Two common modelling techniques, path analysis, and structural equation modelling (SEM) or latent variable modelling, reflect an advanced application of the techniques of regression analysis, allowing researchers to determine possible causal relationships between two or more variables. The following is a simple overview to tantalize you with the possibilities that regression models provide. For further detailed explanations and relevant computer programs like AMOS, the reader is directed to one of a number of advanced research methods texts listed at the end of this chapter.

Path analysis

Path analysis is an extension of the regression model, analysing a complex set of relationships to determine direct and indirect effects between a number of variables. Path analysis models usually depict variables in an ellipse with single arrows indicating causation direction. A regression equation is formulated for each variable identified in the model as a dependent variable. However, although path analysis is very popular, correlational data is still only correlational. Whilst path analysis can tell us which the significant paths are, it cannot inform whether one model is better than another model. Also, it cannot indicate whether a correlation between two variables represents a causal effect.

Developing a path model

The first stage to path analysis is to create a path model (Figure 1) showing the relationships between independent and dependent variables, and possibly intermediary variables. In path analysis, the terms exogenous and endogenous variables are used. Exogenous variables are variables whose cause is outside of the model, whilst endogenous variables are those whose causes are identified within the model.

Arrows are drawn between variables and the direction is dictated by the direction the arrow points, which by convention are usually from left to right. The arrows represent paths between the different variables and the strength of the path is indicated by a path coefficient. These coefficients are obtained from the betas in a multiple regression output and show the direct effect of an independent variable on a dependent variable in the path model. Indirect effects can also be measured and are discussed later.

In a path diagram, it is helpful to draw the arrows so that their widths are proportional to the size of the path coefficients, and it is convention that paths whose coefficients fall below a certain coefficient value or which do not reach a level of significance, are omitted in the output path diagram to save clutter. Additional arrows can be included, pointing to each dependent variable. This signifies the unexplained variance, the variation in that variable due to factors not included in the analysis.

Figure 1 shows a Path Analysis Model (Bogler, 2001) with three exogenous variables (Principal's autocratic-participative decision-making; Principal's transformational leadership; Principal's Transactional Leadership), and two endogenous variables (Teacher's occupation perception; Teacher's job satisfaction) where Teacher's occupation perception functions as an intermediate variable. Path coefficients are included where the coefficient was statistically significant. An example of a direct effect can be seen in the path between 'Teacher's occupation perception' and 'Teacher's job satisfaction'. An indirect effect can be seen between 'Principal's autocratic-participative decision-making' and 'Teacher's job satisfaction' where the path is moderated by the variable 'Teacher's occupation perception'. The curved arrows reflect correlations, and not hypothesized causation, between the exogenous variables: 'Principal's autocratic-participative decision-making' is correlated .14 with 'Principal's transformational leadership'. 'Principal's transformational leadership' correlates –.12 with 'Principal's Transactional Leadership'.

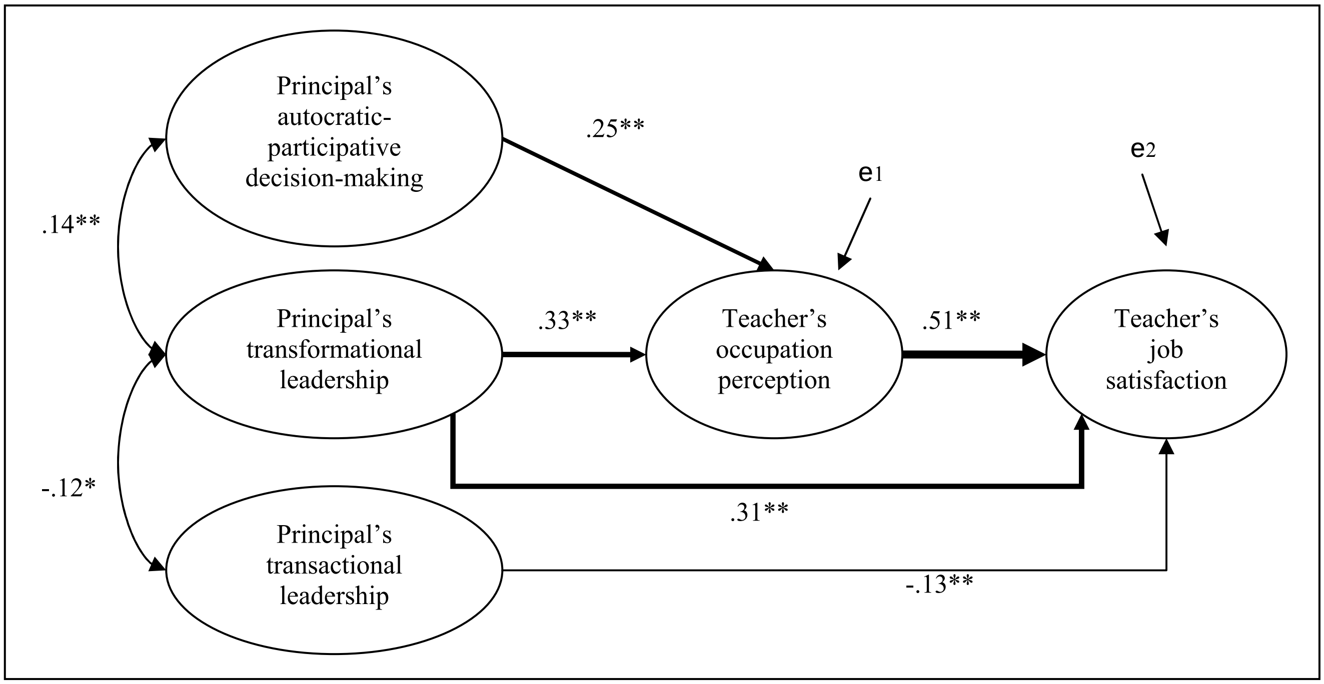

Figure 1 **Path analysis of the relationships between Principal's leadership style, Principal's decision-making strategy, Teacher's occupation perception, and Teacher's job satisfaction (*n* = 706)**
****p* < .001; ***p* < .0001. Bogler R. (2001).**

Path coefficients

To compute path coefficients we need to set up structural equations, which in this case would be:

Equation 1: Teacher's job satisfaction = Teacher's occupation perception + Principal's transactional leadership + Principal's transformational leadership + Principal's autocratic-participative decision-making + e_2.

Equation 2: Teacher's occupation perception = Principal's transactional leadership + Principal's transformational leadership + Principal's autocratic-participative decision-making + e_1.

The model (Figure 1) is specified by these path equations which includes the coefficients for the independent variables and where *e* is the error or unexplained variance term. Since you should now have a full understanding of the procedures to doing regression analysis, obtaining the path coefficients is easy to do. Simply run two regression analyses, with 'Teacher's job satisfaction' and 'Teacher's occupation perception', being the dependent variables and using the independent variables as specified in the equations. Although omitted in this instance, values of *e* are calculated as the square root of 1-R^2 (not the adjusted R^2) from the regression equation for the corresponding dependent variable.

Identifying possible causal effects

Having obtained path coefficients, you are now able to measure a number of effects, including:

- Total Causal Effect: The total of Direct and Indirect effects combined. This indicates the total effect on the dependent variable that could be identified in this model.
- Direct Effects: These are the path coefficients for each variable that is directly related to the dependent variable.
- Indirect Effects: Is the total causal effect minus the direct effect. This allows you to measure the effect of the intervening variables.

Table 1 **Teacher’s job satisfaction**

<table>
<tr><th>Type of Effect</th><th>Compound paths to be computed</th><th>Path coefficients to be multiplied</th><th>Compound Path scores</th></tr>
<tr><td rowspan="8">Indirect effects</td><td>Principal’s autocratic-participative decision-making * Teacher’s occupation perception * Teacher’s job satisfaction</td><td>.25 * .51 =</td><td>.13</td></tr>
<tr><td>Principal’s transformational leadership * Teacher’s occupation perception * Teacher’s job satisfaction</td><td>.33 * .51 =</td><td>.17</td></tr>
<tr><td>Principal’s Transactional Leadership * Principal’s transformational leadership * Teacher’s occupation perception * Teacher’s job satisfaction</td><td>-.12 * .33 * .51* =</td><td>-.02</td></tr>
<tr><td>Principal’s Transactional Leadership * Principal’s transformational leadership * Teacher’s job satisfaction</td><td>-.12 * .33 =</td><td>-.07</td></tr>
<tr><td>Principal’s transformational leadership * Principal’s transactional leadership * Teacher’s job satisfaction</td><td>-.12 * -.13 =</td><td>-.02</td></tr>
<tr><td>Principal’s transformational leadership * Principal’s autocratic -participative decision-making * Teacher’s occupation perception * Teacher’s job satisfaction</td><td>.14 * .25 * .51 =</td><td>.02</td></tr>
<tr><td>Principal’s autocratic-participative decision-making * Principal’s transformational leadership * Teacher’s occupation perception * Teacher’s job satisfaction</td><td>.14 * .33 * .51 =</td><td>.02</td></tr>
<tr><td>Principal’s autocratic-participative decision-making * Principal’s transformational leadership * Teacher’s job satisfaction</td><td>.14 * .31 =</td><td>.04</td></tr>
<tr><td rowspan="3">Direct Effects</td><td colspan="2">Teacher’s occupation perception</td><td>.51</td></tr>
<tr><td colspan="2">Principal’s transactional leadership</td><td>-.13</td></tr>
<tr><td colspan="2">Principal’s transformational leadership</td><td>.31</td></tr>
<tr><td colspan="4">TOTALS</td></tr>
<tr><td colspan="3">Indirect effects total</td><td>.27</td></tr>
<tr><td colspan="3">Direct effects total</td><td>.69</td></tr>
<tr><td colspan="3">Total causal effect</td><td>.96</td></tr>
<tr><td colspan="2" rowspan="3">Principal’s autocratic-participative decision-making</td><td>Indirect effect</td><td>.19</td></tr>
<tr><td>Direct effect</td><td>0</td></tr>
<tr><td>Total effect</td><td>.19</td></tr>
<tr><td colspan="2" rowspan="3">Principal’s transformational leadership</td><td>Indirect effect</td><td>.17</td></tr>
<tr><td>Direct effect</td><td>.31</td></tr>
<tr><td>Total effect</td><td>.48</td></tr>
<tr><td colspan="2" rowspan="3">Principal’s transactional leadership</td><td>Indirect effect</td><td>-.09</td></tr>
<tr><td>Direct effect</td><td>-.13</td></tr>
<tr><td>Total effect</td><td>-.22</td></tr>
</table>

Table 1 indicates the indirect and direct effects of all the variables identified in the model in Figure 1, on the dependent variable, 'Teacher's job satisfaction'. All possible combinations of indirect effects are computed first, followed by the direct effects of 'Teacher's occupation perception', 'Principal's transactional leadership', and 'Principal's transformational leadership'. The totals for direct and indirect indicate that the direct effect of the independent variables is greater than the indirect. The last three rows indicate the direct, indirect and total effect for each of the three exogenous factors. These indicate that whilst 'Principal's autocratic-participative decision-making' has no direct effect it does have an indirect effect on 'Teacher's job satisfaction'. 'Transactional leadership' appears to have negative direct and indirect effects on 'Teacher's job satisfaction'. 'Transformational leadership' appears to have both direct and indirect effects on 'Teacher's job satisfaction'.

Compound paths

The value of the compound paths listed in Table 1 is simply the product of the path coefficients and these scores allow you to determine the strengths of the different paths. For example, consider the simple three-variable compound path where 'Principal's autocratic-participative decision-making' determines 'Teacher's occupation perception' which then determines 'Teacher's job satisfaction'. The regression coefficient of 'Teacher's occupation perception' on 'Principal's autocratic-participative decision-making' is .25: for each point increase on a 4-point scale measuring 'Principal's autocratic-participative decision-making' – where a high score indicates a participative decision-making climate, 'Teacher's occupation perception' goes up .25 of a point. As the regression coefficient of 'Teacher's job satisfaction' on 'Teacher's occupation perception' is .51: for every point increase on a 5-point scale measuring 'Teacher's occupation perception', 'Teacher's job satisfaction' goes up .51 of a point on a 7-point scale. Thus if 'Principal's autocratic-participative decision-making' goes up one scale point, 'Teacher's occupation perception' goes up .25 of a point and 'Teacher's job satisfaction' goes up .13 points (arrived at by multiplying the coefficients between 'Principal's autocratic-participative decision-making' and 'Teacher's occupation perception' and 'Teacher's occupation perception' and 'Teacher's job satisfaction' which in this case is .25 * .51).

Another use for these paths is to allow you to determine the specific amount of effect of 'Teacher's occupation perception' on 'Teacher's job satisfaction', which is reported to be .51 (Figure 1). What this coefficient fails to explain is the amount of this direct effect that can be contributed to the five compound paths whereby 'Teacher's occupation perception' acts as an intermediate variable. By subtracting all indirect effects which pass through 'Teacher's occupation perception' as detailed in Table 1 it is possible to identify that 'Teacher's occupation perception' directly contributes to .19 of the .51 coefficient reported in Figure 1.

However, this process does not really allow you to make an assessment about how well the model fits the data. Instead a causal analysis technique known as Structural Equation Modelling (SEM)/Latent Variable Modelling must be used, and are usually performed by specialized statistics programmes like AMOS and LISREL. These programmes report a number of tests including a goodness of fit chi-square test, and other best model fit tests, which assess the fit of your data by comparing the estimated correlation matrix with the observed correlation matrix. An example of SEM is included below.

Assumptions and limitations of path analysis

Path analysis can evaluate causal hypotheses but it cannot establish the direction of causality. Whilst the direction of arrows reflect hypotheses about causation, path analysis really only allows us to observe which of two or more theory driven models are most consistent with the patterns reported in our data. The following are also important issues:

- The relationships between variables must be linear.
- There should be no interaction effects.
- The relationships in the model need to be appropriate for multiple regression analysis in order that independent and intervening variables function as dependent variables in multiple regression analyses. Interval level data should be used to estimate path parameters, though it is common to generate dummy variables for dichotomies and ordinal data.
- The model should not be recursive so all the arrows flow one way, with no feedback looping.

- A large enough sample size is necessary to assess significance. Common suggestions include 10–20 times as many subjects as variables (Kline, 2005).

Path analysis is useful when we already have a clear hypothesis to test, or a small number of hypotheses all of which can be represented within a single path diagram. Path analysis is limited in recursive models where there is a feedback loop between a dependent variable which in turn influences an independent variable. For example, it is commonly assumed that perceptions of organizational stress determine employee well-being whereby a highly stressful work environment can lead to higher levels of physical and emotional stress. Conversely, it could be proposed that an employee's well-being also determines their perception of the workplace. It is well established that people who report negative feelings are more likely to perceive the world around them in a negative way. In this instance Path analysis would need to establish two models to test the non-recursive nature of the topic. This is something which SEM has attempted to address.

Structural equation modelling/Latent variable modelling

Whilst structural equation modelling (SEM) shares some similarities with traditional path analysis in the modelling of correlated independents, measurement error, and one or more dependent variables, SEM is a more powerful technique in that it accounts for the modelling of interactions, non-linearities, correlated error terms, and multiple latent and dependent independents which are measured by multiple indicators.

Generally, SEM is used to confirm a particular model rather than explore a range of possible best-fitting models. For example, a strictly confirmatory approach involves determining if the data is consistent with the model specified by the researcher, or with alternative models would seek to test two or more causal models to determine which has the best fit according to several different measures of goodness of fit. However, as with path analysis, SEM does not identify the direction of causal associations in models. Such inferences must be drawn from theory and the researcher's personal judgement.

SEMs can include two kinds of variables: observed and latent. Observed variables have data, like the numeric responses to a rating scale item on a questionnaire such as gender or height. Observed variables in SEMs are also usually continuous. Latent variables, e.g. household income, brand attitudes, customer satisfaction, perceived value, repurchase intentions and perceived quality, are not directly observed, but you still want to investigate them. To observe latent variables, you must build models to express latent variables in terms of observed variables, e.g. questionnaire items. The latent variables in SEMs are continuous variables and can, in theory, have an infinite number of values. SEM expresses linear relationships between variables. This is also how variables are usually related in regression and factor analysis models – variables are expressed as weighted linear combinations of other variables.

Variables that depend on other variables are called 'dependent'. Variables that do not depend on other variables in a model are called 'independent.' When using SEMs, these variables are also called 'endogenous' and 'exogenous,' respectively. Consider the linear relationship expressed by Equation 1. This equation says that perceived value for case 'i' is the sum of the quality for 'i' multiplied by the coefficient 'a', cost for 'i' multiplied by the coefficient 'b', plus an 'error'. The error term represents that part of perceived value for case 'i' that is not captured by its linear dependence on quality and cost. When combined with some assumptions, the equation describes a model of value that may depend on quality and cost.

$$[\text{Perceived value of good}_i = a(\text{quality}_i) + b(\text{cost}_i) + \text{error}_i] \qquad \text{(Eq: 1)}$$

This equation looks like a regression equation, without an intercept (constant) term on the right-hand side. The coefficients 'a' and 'b' represent the regression coefficients. 'Value', 'quality' and 'cost' are observed variables. 'Error' is the difference between the observed and predicted values for each of the cases. Or, you may see the equation as describing a factor model in which the observed variable called value 'loads' on two factors called 'quality' and 'cost'. Quality and cost are latent variables. When we fit a model like the model into an equation as above to a data set, we're trying to pick estimates for coefficients 'a' and 'b' that minimize some function of the errors across observations, given some assumptions about these errors. The model assumes that all cases in the data set have the same values for 'a' and 'b'. They are fixed in the population.

Introduction to SEM path diagrams

Figure 2 shows a clear way of describing the model in the equation. Figure 2 is a path diagram. You can draw path diagrams quite easily using the graphics tools in Amos. Amos also generates the necessary equation statements to fit the models you draw. Observed variables are drawn as boxes, latent variables are drawn as circles or ellipses. Note, the error term in the path diagram is drawn as latent – errors are estimated, not measured directly. When one variable is believed to 'cause' another variable, the relationship between the variables is shown as a directed or one-headed arrow, from cause to effect. Whether one variable 'causes' another is an assumption that you make, not something the data can tell you. For each arrow, there may be an estimated loading or weight, like the coefficients 'a' and 'b' in Figure 2.

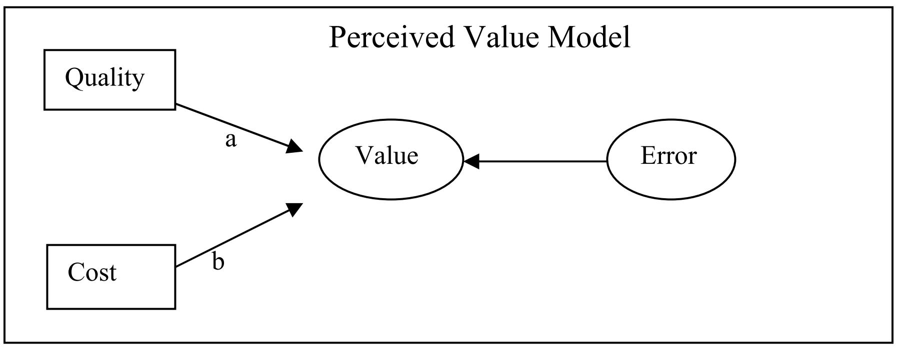

Figure 2 **A path diagram for the value model.**

Sometimes covariation between two variables needs to be included in a SEM. This kind of undirected relationship is shown as a curved, two-headed arrow connecting the variables. Weights constrained to a particular value are often labelled with that value.

SEMs may also include one or more linear regression equations that describe how some variables depend on others. These are called structural equations. The collection of them is sometimes called the structural equations model, or the structural model in an SEM. The coefficients describing how dependent variables depend on independent variables are sometimes called path coefficients.

You can use latent variables in structural equations as dependent or independent variables. When you use a latent variable in a SEM, it is usually modelled using two or more observed variables called 'indicator' variables. For example, you want to model brand loyalty as a latent variable. You ask customers to make quantitative judgements about their use of a brand, their intentions to continue using the brand, and their willingness to recommend the brand to others. You could then use the responses on these indicator variables to model loyalty as a single latent variable. How each indicator variable related to loyalty would be expressed as a factor loading.

As you might imagine, once you've put together some structural equations and some measurement models, you can get a much more complicated model than those examples here. Remember, these models are built using simple parts, even though they can get complex. You'll find path diagrams are an effective way of summarizing even very extensive SEMs.

Another example of SEM is McGowan, *et al*.'s (2006) attempt to identify the causal relationship between an employee's appraisal of stressful events, the types of coping strategies used, the level and type of stress experienced, and a measure of satisfaction with the outcome of the coping strategy employed.

Figure 3 shows that the relationship between questionnaire items (Manifest Variables – represented by squares) and the factors (Latent Variables – represented by circles) they purport to measure are of moderate to strong correlations. Eight items comprised the Cognitive Appraisal Scale (CAS), with four questions each related to threat and challenge appraisals – within the context of this study, question 4 did not load onto either of these two factors. The WCQ Questionnaire assessed coping strategies used by participants to manage stressful events and, 66 items assess eight forms of coping: planful problem solving (PPS), positive reappraisal (PR), seeking social support (SS), confrontive coping (C), escape avoidance (EA), distancing (D), self-controlling (SC) and accepting responsibility (AR). Four of each of the forms of coping strategies are loaded onto two Factors task-focused and emotion-focused coping. The Job Related Affective Wellbeing Scale (JAWS) assessed participants' emotional reactions to their work. Participants rated their recent experience of 30 different emotions and covered two dimensions: positive/negative affect and high/low arousal. Although this allows for four quadrants: negative affect/low arousal, negative affect/high arousal, positive affect/low arousal, and positive affect/high arousal, for this study only two quadrants were measured: 'positive affect/high arousal' (eustress/positive stress); and 'negative affect/high arousal' (distress), with five questions each loaded onto these two factors.

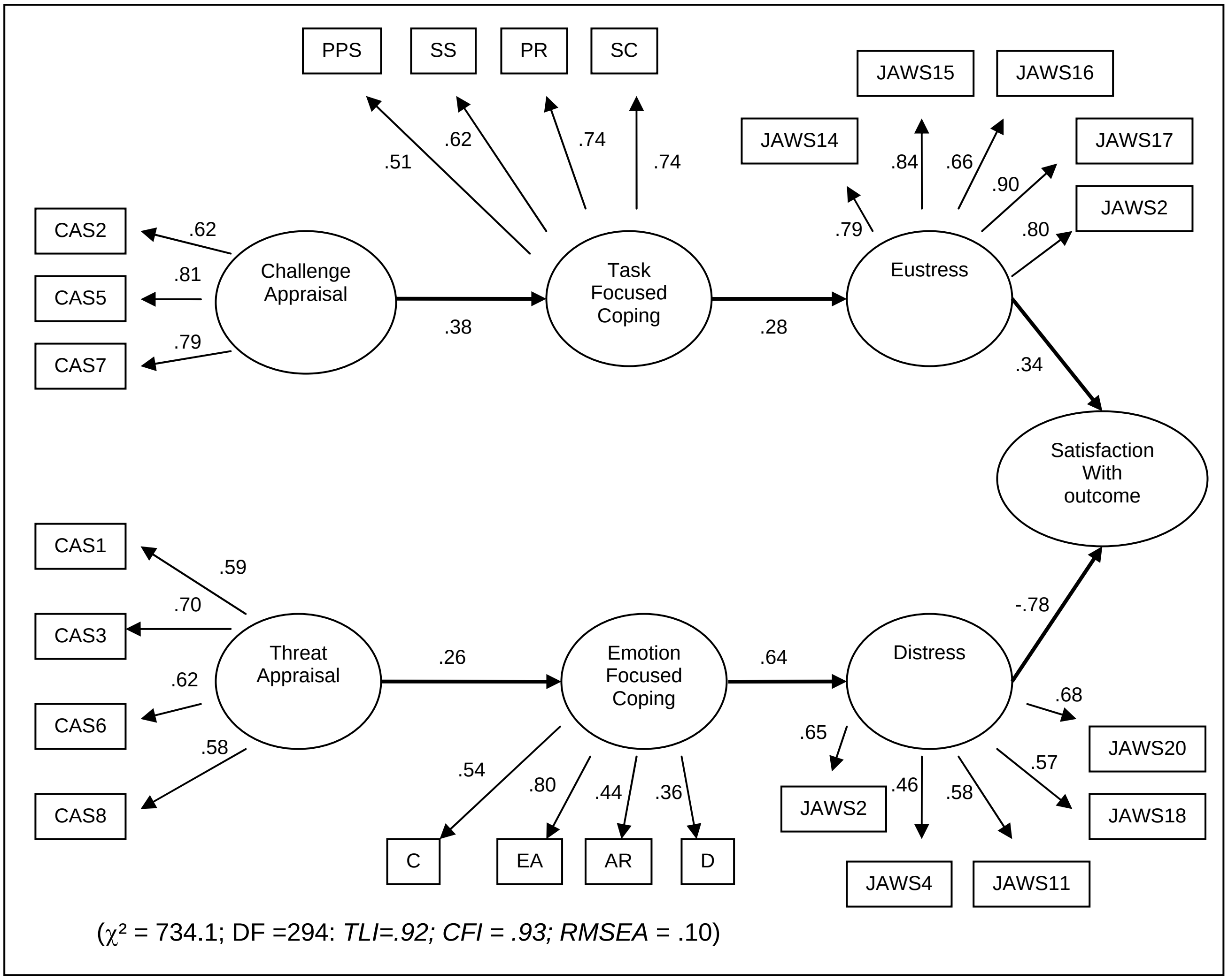

Figure 3 **Full structural model** (McGowan *et al*, 2006).

The SEM also includes a path analysis between the latent variables which demonstrates the proposed causal relationships that exist between these variables which are indicated in Figure 3. Although researchers argue about just how well these models identify the directions of these associations, as with path analysis, a strength with the SEM is the assessment of an overall fit of the data and a significance test which will measure how accurate the data actually fits the proposed model. These can be seen in the statistical analysis reported by McGowan *et al*. (2006) which were (χ^2 = 734.1; DF =294: *TLI=.92; CFI = .93; RMSEA* = .10). You may remember that a large chi square (χ^2 = 734.1) indicates that there is a large difference between Expected and Observed scores, rejecting the null hypothesis that there is no significant difference between Expected and Observed scores. As a brief explanation to the other statistics – since you are unlikely to have seen them before let alone understand them – the TLI and CFI were above .90 a figure which indicates a good fit, whilst RMSEA scores of .10 to .08 indicate an adequate fit and McGowan *et al*. (2006) therefore were able to support their model.

References

Bogler R. (2001) 'The influence of leadership style on teacher job satisfaction', *Educational Administration Quarterly* 37: 662.

Kline, R. (2005) *Principles and Practice of Structural Equation Modelling* (2nd ed). New York : Guilford Press. (A very readable introduction to the subject, with good coverage of assumptions and SEM's relation to underlying regression, factor, and other techniques.)

McGowan, J., Gardner, D. and Fletcher, R. (2006) 'Positive and negative affective outcomes of occupational stress' *New Zealand Journal of Psychology* 35, 2: 95 – 102.