CHAPTER 6

THE REGRESSION MODEL

6.0 Introduction

As stated in Chapter 5, the regression model that emerged from the findings will be discussed in this chapter. The procedure to formulate a regression model that incorporates representational competence as the criterion variable, with prior knowledge, developmental level, working memory capacity, and learning orientations as the predictor variables involves a six-stage model-building framework. These stages are: (i) specifying objectives of the multiple regression, (ii) establishing the research design of the multiple regression analysis, (iii) assessing the assumptions in multiple regression analysis, (iv) estimating the regression model and assessing overall model fit, (v) interpreting the regression variate, and (vi) validating the regression model. The following sub-sections present the detailed procedure of each of these stages. The chapter concludes by linking the regression model to theory.

6.1 Objectives of the Multiple Regression

Multiple regression analysis is a multivariate statistical technique used to examine the relationship between a single dependent (criterion) variable, and a set of independent (predictor) variables.

Application of multiple regression falls into two broad classes of research problems: prediction and explanation. Prediction involves the extent to which the regression variate (one or more independent variables) can predict the dependent variable. Explanation examines the regression coefficients (their magnitude, sign, and statistical significance) for each independent variable and attempts to develop a

theoretical reason for the effects of the independent variables. An application of multiple regression can address either or both types of research problems. In this study, multiple regression analysis was employed as the statistical technique to predict representational competence in chemistry. In addition, factors affecting representational competence would also be identified and explained.

6.2 Research Design of the Multiple Regression Analysis

As multiple regression is a dependence technique, the variables involved must be divided into dependent and independent variables, and both types of variables must be metric. To apply the regression procedure, representational competence had been selected as the dependent variable (Y), to be predicted by independent variables affecting representational competence. The five independent variables were: (i) understanding of chemical concepts or prior knowledge I (X_1), (ii) developmental level (X_2), (iii) understanding of chemical representations or prior knowledge II (X_3), (iv) learning orientation (X_4), and (v) working memory capacity (X_5). In this study, representational competence was measured using the Test of Representational Competence (TRC) while understanding of chemical concepts, understanding of chemical representations, developmental level, learning orientations, and working memory capacity were assessed by the Test of Chemical Concepts (TCC), Test of Chemical Representations (TCR), Classroom Test of Scientific Reasoning (CTSR), Learning Approach Questionnaire (LAQ), and Digit Span Backwards Test (DSBT) respectively. Hence, TRCt score became a measure of representational competence (Y) while the independent variables X_1, X_2, X_3, X_4, and X_5 were measured by the TCCt score, CTSR score, TCRt score, LAQ score, and DSBT score respectively. A total of six instruments were used to collect data for the six variables in this study. For the purpose of model-building, only subjects who

were administered all the six instruments (n=192) were included in the multiple regression analysis. See Chapter 5 - Table 5.1.

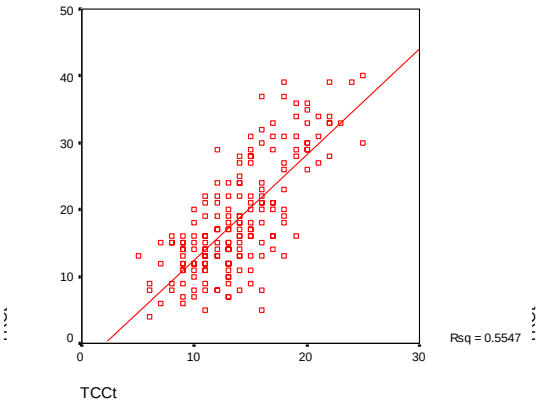
In terms of statistical power and sample size, Table 4.5 (Hair et al., 2006, p.174) shows the minimum percentage of variance explained (R^2) that can be found statistically significant with a power of 0.80 for varying numbers of independent variables and sample sizes. At a significant level (α) of 0.01, R^2 values of 10% and above could be detected with 5 independent variables and a sample size of 200. If the significant level is relaxed to 0.05, then the analysis could identify relationships explaining about 8% of the variance. In terms of generalizability and sample size, the sample of 192 observations also meets the minimum ratio of observations to independent variables (5:1), with an actual ratio of approximately 38:1 (192 observations with 5 independent variables). However, if a stepwise procedure is employed, the recommended level increases to 50:1 (Hair et al., 2006, p.175). Hence, the sample size of n=192 (approximately 200) with 5 independent variables should meet the criteria of both statistical power and generalizability to employ multiple regression analysis.

6.3 Testing for Statistical Assumptions in Multiple Regression Analysis

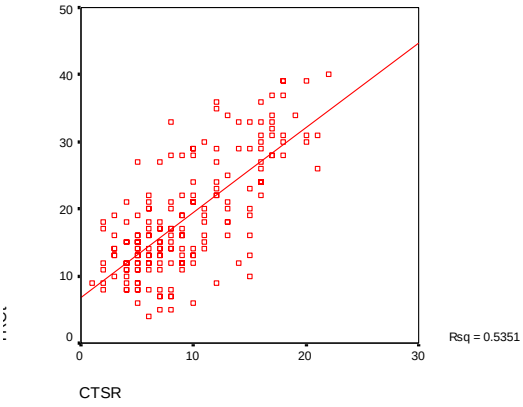
Several statistical assumptions about the relationships between the independent and dependent variables that affect the statistical procedure used for multiple regression must be made. Useful insight is gained in examining the individual variables. Analyses to examine the variate and its relationship with the dependent variable for meeting the assumptions of multiple regression must be performed after the regression model has been estimated. In this section, the three basic assumptions to be addressed for the individual variables are linearity, homoscedasticity, and normality.

6.3.1 Linearity

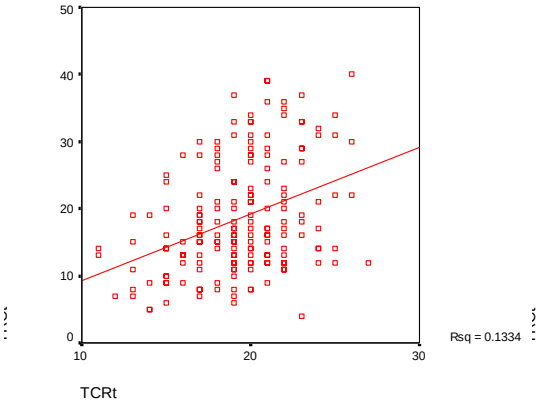
To assess linearity of the data through visual inspection, scatter plots of the individual variables were obtained (Figures 6.1a to 6.1e).



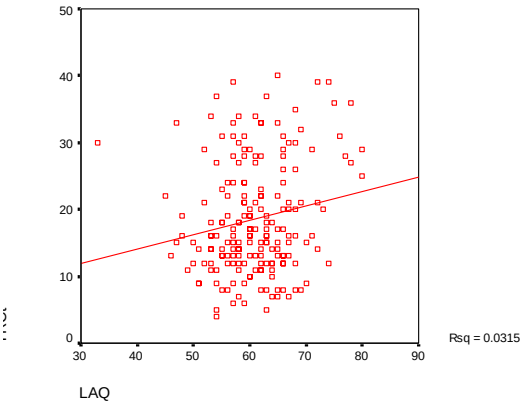
(a) TRCt scores vs TCCt scores



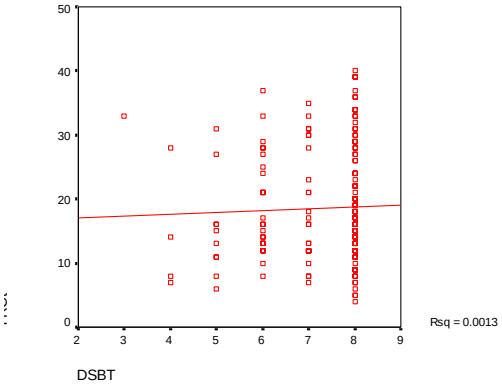
(b) TRCt scores vs CTSR scores



(c) TRCt scores vs TCRt scores



(d) TRCt scores vs LAQ scores



(e) TRCt scores vs DSBT scores

Figure 6.1: Scatter plots of the independent variables

Examination of the scatter plots of the individual variables did not reveal any apparent non-linear patterns or relationships between the dependent variable and independent variables. Thus, transformations are not necessary.

6.3.2 Homoscedasticity

Homoscedasticity refers to the assumptions that dependent variable(s) exhibit equal levels of variance across the range of predictor variable(s). According to Hair et al., (2006), homoscedasticity is desirable because the variance of the dependent variable being explained in the dependence relationship should not be concentrated in only a limited range of the independent values.

Tests for homoscedasticity of two metric variables in methods such as multiple regression are best examined graphically, based on the dispersion of the dependent variable across the values of the independent variables. In scatter plots, departures from an equal dispersion are shown by such shapes as cones (small dispersion at one side of the graph, large dispersion at the opposite side) or diamonds (a large number of points at the centre of the distribution). See Figures 6.1a to 6.1e.

Test for homoscedasticity found that only one independent variable (DSBT scores) violated this assumption. Another independent variable (LAQ scores) had minimal violation. However, no corrective action was needed as a quick check of multiple regression analysis using SPSS showed that these two independent variables LAQ scores (X_4) and DSBT scores (X_5) were not statistically significant contributors to the regression model.

6.3.3 Normality

Normality is required to use the ***F*** and ***t*** statistics. Hence, the most fundamental assumption in multivariate analysis such as multiple regression analyses

is normality. In this chapter, both the graphical plots and statistical tests were used to assess the actual degree of departure from normality.

6.3.3.1 Graphical analysis of normality

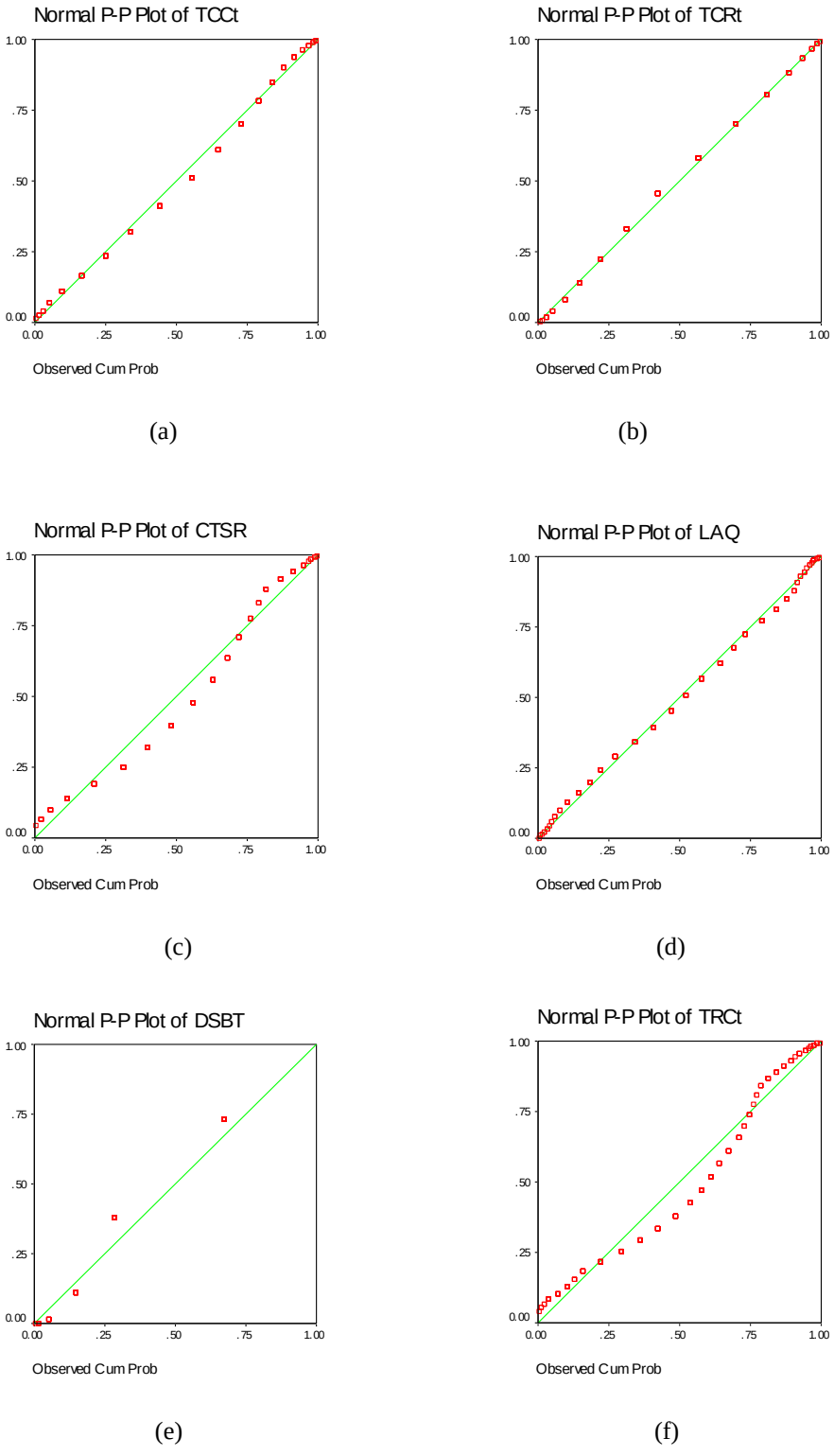


Figure 6.2: Normal probability plots

A reliable approach of graphical analysis is the normal probability plot, which compares the cumulative distribution of actual data values with the cumulative distribution of a normal distribution. The normal distribution forms a straight diagonal line, and the plotted data values are compared with the diagonal. If a distribution is normal, the line representing the actual data distribution closely follows the diagonal. See Figures 6.2a to 6.2f.

6.3.3.2 Statistical tests of normality

(i) Statistical tests of normality assess the degree to which the skewness and peakedness of the distribution vary from the normal distribution where $Z_{skewness}$=skewness/√6/N and $Z_{kurtosis}$=kurtosis/√24/N. If either calculated z value exceeds the specified critical value, then the distribution is non-normal in terms of that characteristic. Critical values are +/-2.58 for α=0.01 and +/-1.96 for α=0.05.

(ii) Specific statistical tests for normality - the modified Kolmogorov-Smirnov test calculates the level of significance for the differences from a normal distribution. The Kolmogorov-Smirnov statistic with a Lilliefors significance level for testing normality is produced with the normal probability plot. If the significance level is >0.05, normality is assumed.

When viewing the shape characteristics, significant deviations were found for skewness (X_2, X_5 and Y) and kurtosis (X_5). The normal probability plots can also be used to identify the shape of the distribution. Figures 6.2(c), (e) and (f) contain the normal probability plots for the three variables found to have non-normal distribution. These three variables (X_2, X_5 and Y) were also found to violate the statistical tests. Of the five independent variables, only X_2 and X_5 show any deviation from normality in the overall normality tests (See Table 6.1). Overall, departures from normality are not so extreme in any of the variables.

Table 6.1
Distributional characteristics and testing for normality

Variables	Shape descriptors				Statistical test for normality (the Kolmogorov-Smirmov test)		Normal Probability Plots
	Skewness		Kurtosis				
Independent Variables	Statistic	Z value	Statistic	Z value	Statistic	Significance	Description of distribution
TCCt (X_1)	.418	**2.388**	-.069	**-.198**	1.290	**.072**	normal
CTSR (X_2)	.602	3.440	-.577	**-1.653**	1.734	.005	Negative Distribution
TCRt (X_3)	-.203	**-1.160**	-.052	**-.148**	1.461	.028	≈normal
LAQ (X_4)	.003	**0.017**	1.202	3.444	0.747	**.632**	normal
DSBT (X_5)	-1.595	-9.114	1.905	5.458	5.325	.000	Uniform distribution
Dependent Variable TRCt (Y)	.665	3.800	-.432	**-1.238**	1.893	.002	Negative Distribution

6.3.4 Section Summary

The series of graphical and statistical tests directed towards assessing the assumptions underlying the multivariate techniques revealed relatively little in terms of violations of the assumptions. Where violations were detected, they were relatively minor and should not present any serious problems in the course of the data analysis. Although normality can have serious effects in small samples (fewer than 50 cases), the impact effectively diminishes when sample sizes reaches 200 cases or more, as large sample sizes tend to diminish the detrimental effects of non-normality. According to Hair et al., (2006), even analysis with small sample sizes can sometimes withstand small, but significant departures from normality. In this study, the use of multiple regression analysis to more accurately predict the criterion variable of representational competence far outweighs the little violation, and with the relatively large sample size (n=192), the procedure should be attempted.

6.4 Estimating the Regression Model and Assessing Overall Model Fit

To derive the regression equation, the method of estimation must be decided and the number of independent variables to be retained determined. Three basic tasks to be accomplished at this stage are: (i) selecting a method for specifying the regression model to be estimated, (ii) assessing the statistical significance of the overall model in predicting the dependent variable, and (iii) determining whether any of the observations exert an undue influence on the results. The detail procedure for each of these tasks was discussed in the following sub-sections.

6.4.1 Selecting an estimation technique

There are 5 independent variables to choose for inclusion in the regression equation. The set of independent variables can either be exactly specified and the regression model used in a confirmatory approach or estimation technique used to pick and choose among the set of independent variables. Estimation techniques that can be used include: (i) confirmatory specification, (ii) sequential search methods, and (iii) combinatorial approach.

6.4.1.1 Estimating the regression model using sequential search method

In this study, the sequential search method - the stepwise regression, was used for estimating the regression model. The stepwise estimation procedure is designed to develop a regression model with the fewest number of statistically significant independent variables and maximum predictive accuracy. The ability of the stepwise method to add and delete makes it the preferred method of estimation technique. Moreover, the stepwise procedure maximizes the incremental explained variance at each step of model making. Although with computerised estimation procedures using SPSS, multiple regression analysis can be done almost instantly,

the procedure of the stepwise regression analysis needs to be discussed with reference to the findings of this study.

Procedure of the stepwise regression analysis

Stepwise estimation: Step 1 - Entering the 1st variable, X_1

Start with the simple regression model by selecting the one independent variable that is the most highly correlated with the independent variable. The equation would be $Y=b_0 + b_1X_1$. Table 6.2 displays all the correlations among the 5 independent variables and the correlations with the dependent variable (Y), which is representational competence in this study.

Table 6.2
Correlation matrix (n=192)

	Y	X_1	X_2	X_3	X_4	X_5
Dependent Variable						
Y TRCt scores						
Independent Variables						
X_1 TCCt score	0.745***	1.000				
X_2 CTSR score	0.731***	0.575***	1.000			
X_3 TCRt score	0.365***	0.293***	0.181**	1.000		
X_4 LAQ score	0.178**	0.136*	0.087	0.100	1.000	
X_5 DSBT score	0.036	0.010	0.123*	0.006	-0.147*	1.000

*** correlation significant at p<0.001
** correlation significant at p<0.01
* correlation significant at p<0.05

Examination of the correlation matrix (see Table 6.2) reveals that TCCt scores (X_1) has the highest bivariate correlation of 0.745 with the dependent variable. The first step is to build a regression equation using just this single independent variable. Table 6.3 shows the regression results of this first step.

Table 6.3
Step 1 of the multiple regression analysis

Step 1 – Variable entered: TCCt score (X_1)

Multiple R	.745
Coefficient of Determination (R^2)	.555
Adjusted R^2	.552
Standard error of the estimate	5.643

Analysis of variance (ANOVA)

	Sum of Squares	**df**	Mean Square	*F*	Sig.
Regression	7535.787	1	7535.787	236.648	.000[a]
Residual	6050.332	190	31.844		
Total	13586.120	191			

Variables entered into the regression model

	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics	
Variable Entered	B	Std. Error	Beta	*t*	Sig.	Zero-order	Partial	Part	Tolerance	VIF
(Const.)	-3.286	1.480		-2.221	.028	-	-	-	-	-
X1 TCCt scores	1.576	.102	.745	15.383	.000	.745	.745	.745	1.000	1.000

Variables not entered into the regression model

	Beta In	Statistical Significance		Partial Correlation	Collinearity Statistics	
		t	Sig.		Tolerance	VIF
X2 CTSR scores	.453[a]	9.186	.000	.556	.670	1.493
X3 TCRt scores	.161[a]	3.257	.001	.231	.914	1.094
X4 LAQ scores	.078[a]	1.603	.111	.116	.982	1.019
X5 DSBT scores	.029[a]	.597	.551	.043	1.000	1.000

a Predictors in the Model: (Constant), TCCt

A discussion of the overall model fit as well as the first step of the model estimation is given below:

(i) The multiple R is the same as the bivariate correlation (0.745) because in the first step of the stepwise estimation, the equation contains one variable.

(ii) The coefficient of determination or R^2 which is ($0.745^2 = 0.555$ or 55.5%) indicates the percentage of total variation of representational competence (Y), explained by the regression model consisting of TCCt score (X_1).

(iii) The standard error of the estimate is a measure of the accuracy of predictions. It is a measure to assess the absolute size of the prediction error.

(iv) ANOVA and *F* ratio – The ANOVA analysis provides the statistical test for the overall model fit in terms of the *F* ratio. The total sum of squares (7535.787 + 6050.332 = 13586.120) is the squared error that would occur if only the mean of Y is used to predict the dependent variable, Y. Using the values of TCCt scores (X_1) to predict the dependent variable (Y) reduces the square error by 55.5% (7525.787/13586.120). This reduction is statistically significant, F (1, 190) = 236.648, $p < 0.001$.

Stepwise estimation: Step 2 - Adding a 2nd variable, X_2

In this study, a 0.10 level is set for dropping variables from the equation. The next step in a stepwise estimation is to check and delete any of the variables in the equation that now fall below the significant threshold, and once done, add the variable with the highest statistically significant partial correlation. From Table 6.3, CTSR score (X_2), with a partial correlation coefficient of 0.556 would be the next independent variable to be entered. The following section provides the details of the newly formed regression model and the issues regarding its overall model fit, the

estimated coefficients, the impact of multicollinearity, and identification of a variable to add in the next step.

Table 6.4
Step 2 of the multiple regression analysis

Step 2 – Variable entered: CTSR score (X_2)

Multiple R	.832[b]
Coefficient of Determination (R^2)	.692
Adjusted R^2	.689
Standard error of the estimate	4.704

Analysis of variance

	Sum of Squares	df	Mean Square	*F*	Sig.
Regression	9403.208	2	4701.604	212.437	.000[b]
Residual	4182.911	189	22.132		
Total	13586.120	191			

Variables entered into the regression model

	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics	
Variable Entered	B	Std. Error	Beta	*t*	Sig.	Zero-order	Partial	Part	Tolerance	VIF
(Const.)	-2.905	1.234		-2.353	.020	-	-	-	-	-
X1	1.025	.104	.484	9.819	.000	.745	.581	.396	.670	1.493
X2	.783	.085	.453	9.186	.000	.731	.556	.371	.670	1.493

Variables not entered into the regression model

	Beta In	Statistical Significance		Partial Correlation	Collinearity Statistics	
		t	Sig.		Tolerance	VIF
X3 TCRt scores	.155[b]	3.787	.000	.266	.914	1.094
X4 LAQ scores	.074[b]	1.822	.070	.132	.981	1.019
X5 DSBT scores	-.025[b]	-.603	.547	-.044	.979	1.021

b Predictors in the Model: (Constant), TCCt, CTSR

Overall model fit

With the addition of X_2, the multiple R and R^2 values have both increased (see Table 6.4). R^2 increased by 0.137 or 13.7%, yielding a total variance explained (R^2) of 0.692 or 69.2%. The adjusted R^2 also increased to 0.689 and the standard error of the estimate decreased from 5.643 to 4.704. Both of these measures demonstrate the improvement in overall model fit.

Estimated coefficients

The regression coefficient for X_2 is 0.783 and the beta coefficient is 0.453 (see Table 6.4). Although not as large as the beta for X_1 (0.484), X_2 still has a substantial impact on the overall regression model. The coefficient is statistically significant at $p<0.001$.

Impact of multicollinearity

Multicollinearity poses a problem. With a tolerance value of 0.670 for both X_1 and X_2, 33% of either variance is explained by the other. Multicollinearity results in substantial change for either the value of b_1 (from 1.576 to 1.025) or the beta value of X_1 (from 0.745 to 0.484) in step 1 of the regression analysis. It further indicates that the variables X_1 and X_2 are moderately correlated, with a correlation coefficient of 0.575 (Table 6.2). However, the *t* values indicate that both X_1 and X_2 are statistically significant predictors of Y.

Identifying variables to add

Since both X_1 and X_2 make significant contributions, neither will be dropped in the stepwise estimation procedure. Looking at the partial correlation for the variables not in the equation in Table 6.4, X_3 has the highest partial correlation (0.266), which is also statistically significant at $p< 0.001$.

Stepwise estimation: Step 3 - A 3rd variable, X_3 is added

The next step in stepwise estimation follows the same pattern of (i) first checking and deleting any variables in the equation falling below the significant threshold and then, (ii) adding the variable with the highest statistically significant partial correlation. The followings section gives the details of the newly formed regression model and the issues regarding its overall model fit, the estimated coefficients, the impact of multicollinearity, and identification of a variable to add in the next step.

Overall model fit

Entering X_3 into the regression equation gives the results shown in Table 6.5. The value of R^2 increased by (0.714 - 0.692 = 0.220) or 2.2%, the adjusted R^2 increased to 0.709, the standard error of the estimate decreased to 4.547. The new variable entered, X_3, makes relatively little contribution to overall model fit.

Estimated coefficients

The addition of X_3 brought a 3rd statistically significant predictor of representational competence into the regression equation. However, the beta coefficient of 0.155 is the lowest among the three predictor variables in the model.

Effects of multicollinearity

Of the three variables in the regression model, the highest tolerance value is for X_3 (0.914), indicating that only 8.6% of variance of X_3 is accounted for by the other two variables, X_1 and X_2. Moreover, with the inclusion of X_3, the tolerance values of X_1 and X_2 have been reduced to 0.633 (from 0.670) for X_1 and to 0.669 (from 0.670) for X_2 in step 2 of the regression analysis. Judging from the change in tolerance values of X_1 and X_2, it could be inferred that the variables X_3 and X_2 are relatively

independent whereas X_3 and X_1 are relatively more correlated to each other. The simple correlations between X_3 and X_2 ($r=0.181$, $p<0.01$) and between X_3 and X_1 ($r=0.293$, $p<0.001$), also supports these inferences (see Table 6.2).

Identifying variable to add

At this stage, only three variables (X_1, X_2, and X_3) have the statistically significant partial correlation necessary for inclusion in the regression equation. By viewing the bivariate correlations of each variable with Y in Table 6.2, it can be seen that X_4 is weakly correlated with Y ($r= 0.178$, $p<0.01$), while the independent variable X_5 had non-significant bivariate correlation ($r=0.036$) with the dependent variable. The partial correlations of both X_4 and X_5 are non-significant at each stage of the regression analysis.

As expected, in subsequent steps in the stepwise analysis, the 4^{th} and 5^{th} variables (X_4 and X_5) are not entered into the regression equation, and neither are any of the variables entered previously removed.

For purposes of conciseness, details of these subsequent steps and the repeating stepwise estimation were omitted. The final regression model shall be the regression model with three variables included (X_1, X_2, and X_3). That is: the model at step 3 of the regression analysis. See Table 6.5.

In this estimation, the probability of *F* to enter and to remove was set at 0.05 and 0.10 respectively. When the stepwise procedure was repeated at a more stringent threshold of 0.01 and 0.05 respectively, the regression results remained unchanged.

Overall model fit

The final model (see Table 6.5) with three independent variables explains almost 71% of the variance of representational competence (Y). The adjusted R^2 of 0.709

indicates no over-fitting of the model and that the results should be generalizable from the perspective of the ratio of observations to independent variables in the equation (192/3 = 64:1) for the final model.

Estimated coefficients

The three regression coefficients, plus the constant, are all statistically significant at $p < 0.001$. See Table 6.5

Impact of multicollinearity

The impact of multicollinearity, in particular between X_1 and X_2, is substantial. With tolerance values of 0.633 and 0.669 for X_1 and X_2 respectively, at least one-third (36.7% and 33.1% respectively) of their variance is accounted for by the other variables in the equation. Although multicollinearity will always affect a variable's contribution to the regression model (p.219, Hair et al., 2006), both X_1 and X_2 are still substantial contributors in the regression model. In contrast, X_3, despite a high tolerance value of 0.914, is a marginal contributor in the regression model ($\beta = 0.155$). See Table 6.5.

The regression model at this stage consists of three independent variables: X_1, X_2, and X_3. Examining the partial correlation of variables not in the equation at this stage, none of the remaining variables have a significant partial correlation at $p < 0.05$ needed for entry. Moreover, all of the variables in the regression model remain statistically significant at $p < 0.001$, avoiding the need to remove a variable in the stepwise process (see Table 6.5). Thus, no more variables are considered for entry or exit and the model is finalized.

Table 6.5
Step 3 of the multiple regression analysis

Step 3 – Variable entered: TCRt score (X_3)

Multiple R	.845[c]
Coefficient of Determination (R^2)	.714
Adjusted R^2	.709
Standard error of the estimate	4.547

Analysis of variance

	Sum of Squares	df	Mean Square	*F*	Sig.
Regression	9699.731	3	3233.244	156.405	.000[c]
Residual	3886.389	188	20.672		
Total	13586.120	191			

Variables entered into the regression model

	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics	
Variable Entered	B	Std. Error	Beta	*t*	Sig.	Zero-order	Partial	Part	Tolerance	VIF
Const.	-9.731	2.161		-4.502	.000					
X1	.933	.104	.441	8.988	.000	.745	.548	.351	.633	1.580
X2	.777	.082	.450	9.440	.000	.731	.567	.368	.669	1.494
X3	.421	.111	.155	3.787	.000	.365	.266	.148	.914	1.094

Variables not entered into the regression model

	Beta In	Statistical Significance		Partial Correlation	Collinearity Statistics	
		t	Sig.		Tolerance	VIF
X4 LAQ scores	.065[c]	1.644	.102	.119	.978	1.023
X5 DSBT scores	-.025[c]	-.626	.532	-.046	.979	1.021

c Predictors in the Model: (Constant), TCCt, CTSR, TCRt

6.4.1.2 Section summary

Table 6.6 provides a step-by-step summary detailing the measures of overall fit for the regression model derived, in predicting representational competence.

Table 6.6
Model summary of stepwise multiple regression

Model Summary[d]

	Overall Model Fit				R^2 Change Statistics				
Step	R	R^2	R^2_{adj}	Std. Error of the Estimate	R^2 Change	*F* Change	df1	df2	Sig. *F* Change
1	.745[a]	.555	.552	5.643	.555	236.648	1	190	.000[a]
2	.832[b]	.692	.689	4.704	.137	212.437	2	189	.000[b]
3	.845[c]	.714	.709	4.547	.022	156.405	3	188	.000[c]

a Predictors: (Constant), TCCt
b Predictors: (Constant), TCCt, CTSR
c Predictors: (Constant), TCCt, CTSR, TCRt
d Dependent Variable: TRCt

From Table 6.6, it could be seen that the first two variables added to the equation made substantial contributions to the overall model fit, with substantive increase in the R^2 and adjusted R^2, while decreasing the standard error of the estimate. It could be inferred that variables X_1 and X_2 are important in assessing overall model fit. With only the first variable (X_1), almost 55% of the variance in representational competence is explained. The second variable (X_2) explains about 14% of the remaining variance, while the third variable, although statistically significant, makes much smaller contribution. X_3 only explains about 2% of the variance.

6.4.2 Assessing the variate for meeting the assumptions of regression analysis

In multiple regression, once the variate is derived, it acts collectively in predicting the dependent variable, which necessitates assessing the assumptions not only for the individual variables but also for the variate itself. Therefore, testing for assumptions must occur not only in the initial phases of the regression, but also after the model has been estimated. Graphical analyses such as partial regression plots, residual plots, and normal probability plots are the most widely used methods of assessing assumptions for the variate. There are also statistical tests that can complement the visual examination of the residual plots.

6.4.2.1 Linearity of the phenomenon

The concept of correlation is based on a linear relationship, thus making it a critical issue in regression analysis. Partial regression plots show the relationship of a single independent variable to the dependent variable, controlling for the effects of all other independent variables. The scatterplot of points depicts the partial correlation between the two variables, with the effects of other independent variables held constant.

Visual examination of the partial regression plots (Figures 6.3a to 6.3e) does not reveal any non-linear relationship between the dependent variable and any of the independent variable. Hence, no corrective action is deemed necessary.

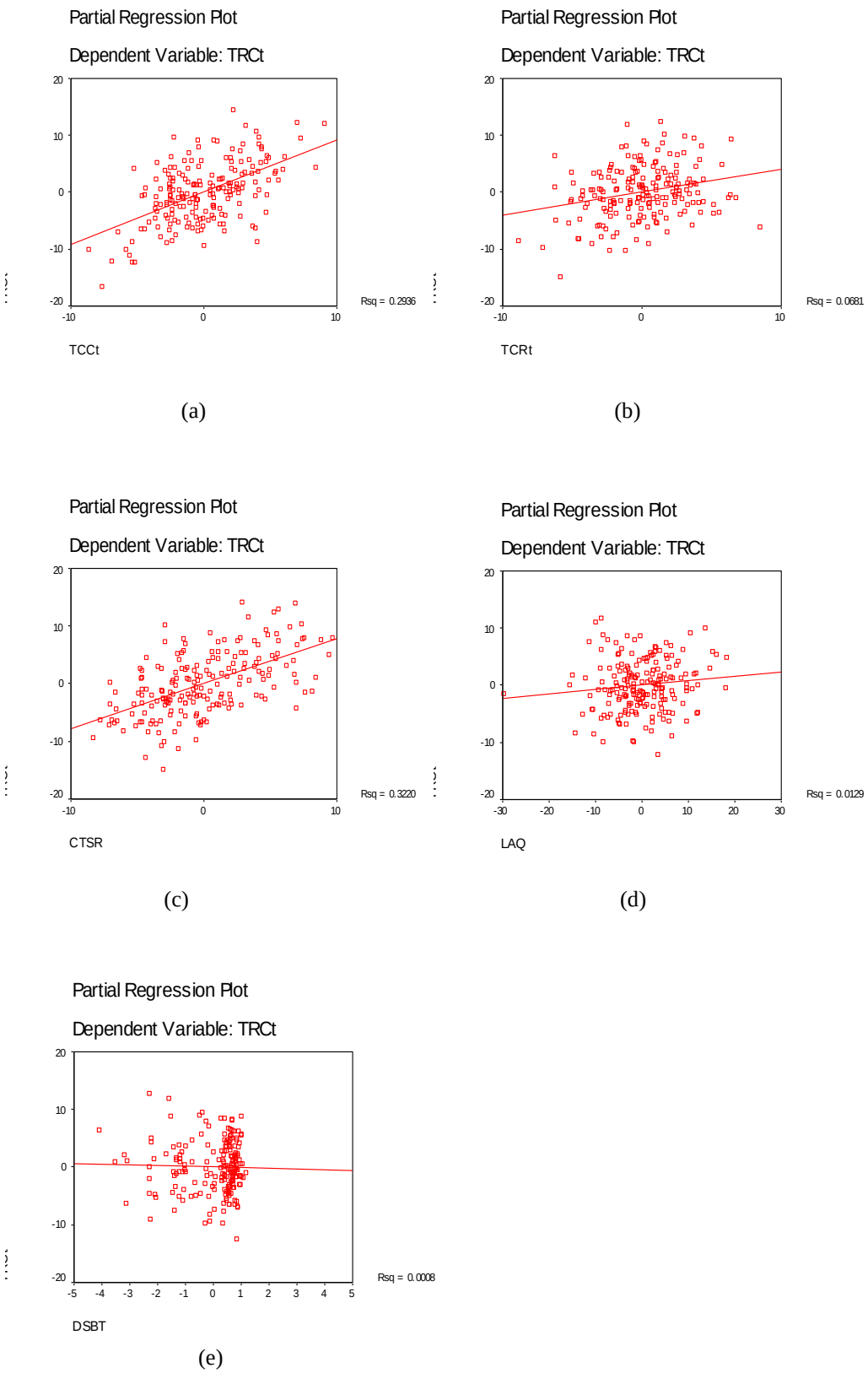


Figure 6.3: Partial regression plots of the independent variables

6.4.2.2 Normality of the error term distribution

The most frequently encountered assumption violation is non-normality of the independent or dependent variables or both. The simplest diagnostic test for normality is a visual check of the histogram (Figure 6.4) that compares the observed data values with a distribution approximating the normal distribution.

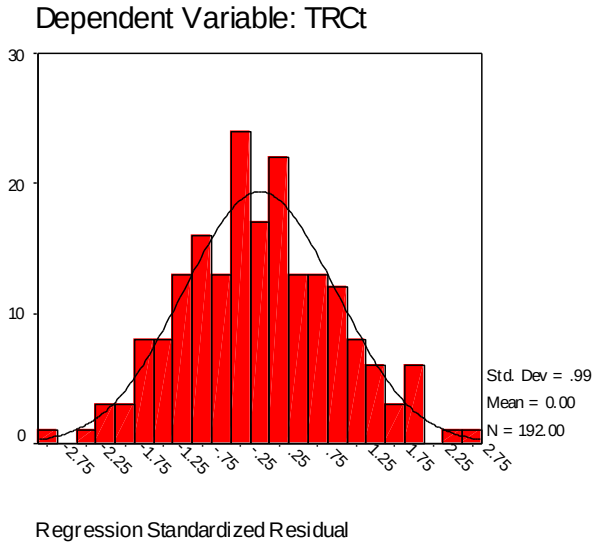


Figure 6.4: Histogram of the dependent variable

A better method is the use of normal probability plots where the standardized residuals are compared with the normal distribution. The normal distribution makes a straight diagonal line, and the plotted residuals are compared with the diagonal. If a distribution is normal, the residual line closely follows the diagonal. See Figure 6.5.

Visual examination of the histogram (Figure 6.4) and the normal probability plot (Figure 6.5) shows little or no violation of normality of the dependent variable.

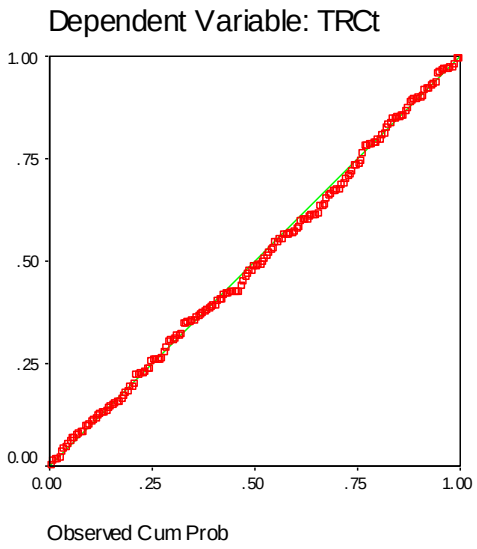


Figure 6.5: Normal probability plot of regression standardised residuals

6.4.3 Identifying unusual observations

To increase the predictive accuracy of the model and the validity of the estimated coefficients, influential observations such as outliers need to be identified. Inspection of the partial regression plots shows only one outlier is detected in Figure 6.3d. However, since learning orientations (measured by the LAQ score) is not a significant contributor to the final regression model, no action was needed.

6.5 Interpreting the Results of Regression or the Regression Variate

After the model estimation is completed, the regression variate needs to be interpreted by assessing the estimated regression coefficients for their explanation of the dependent variables.

6.5.1 Interpreting and using the regression coefficients

Multiple regression provides a means of objectively assessing the degree and character of the relationship between dependent and independent variables by forming the variate of independent variables and then examining the magnitude, sign,

and statistical significance of the regression coefficient for each of the independent variable. In this way, the independent variable, in addition to their collective prediction of the dependent variable, may also be considered for their individual contribution to the variate and its predictions. The regression coefficients for the regression equation with 3 independent variables (X_1, X_2, and X_3), all statistically significant at $p<0.001$, are shown in Table 6.7 below.

Table 6.7
Coefficients of variables in the regression equation

Variables	Regression coefficients			Statistical significance	
	b	Std. error	β	*t*	sig.
[Constant]	-9.731	2.161	-	-4.502	.000
X_1 TCCt score	.933	.104	.441	8.988	.000
X_2 CTSR score	.777	.082	.450	9.440	.000
X_3 TCRt score	.421	.111	.155	3.787	.000

The regression coefficients play two key functions in meeting the objectives of prediction and explanation for any regression analysis. Comparison between regression coefficients allows for a relative assessment of each variable's importance in the regression model.

6.5.1.1 Prediction

Prediction is an integral element in regression analysis. It is important for a regression model to have accurate prediction to support its validity. Independent variable with larger regression coefficient (b) makes a greater contribution to the predicted value.

All the three independent variables, including the constant, were statistically significant at $p < 0.001$, suggesting that they all make a substantive contribution to the prediction. In the predictive process, apart from predicting the value of Y (representational competence), the constant provides no insight for interpretation. In assessing regression coefficients, the sign of the regression coefficients indicates the relationship (+/-) between the independent and dependent variables. Besides the constant term, all the other variables have positive coefficients, suggesting that higher values of these independent variables (X_1, X_2, and X_3) lead to higher value of Y.

The regression equation derived in this study in the form

$Y = b_0 + b_1X_1 + b_2 X_2 + b_3X_3$, that is: $Y = -9.731 + 0.933X_1 + 0.777X_2 + 0.421X_3$, can be used to estimate the representational competence of a student. The expected value of Y could be calculated if values for X_1, X_2, and X_3 were available. For example: The representational competence of a student with the set of data $X_1=24$, $X_2=20$, $X_3=22$ could be calculated by substituting the values for X_1, X_2, and X_3 into the regression equation and calculate the predicted value of Y.

$Y= -9.731 + 0.933(24) + 0.777(20) + 0.421(22) = 37.463$

The regression equation would predict that this student would have a TRCt score of 37.46.

6.5.1.2 Explanation

For explanatory purposes, the regression coefficients become indicators of the relative impact and importance of the independent variables in their relationship with the dependent variable. Insight into the relationship between independent and dependent variables is gained by examining the relative contributions of each independent variable.

In order to use the regression coefficients for explanatory purposes, all the independent variables must be comparable in both scale and variability. These objectives can be achieved by standardizing the regression coefficients to give the standardized or beta coefficients. Beta coefficients reflect the relative impact on the dependent variable of a change in one standard deviation in each independent variable. It allows for a direct comparison between coefficients as to their relative explanatory power of the independent variables.

Interpretation using the regression versus the beta coefficients (β) yields substantially different results. For example, the regression coefficients (b) indicate that X_1 (0.933) is the most important while X_3 (0.421) is only marginally important. The beta coefficients (β), however, show a different set of results. While X_3 (0.155) remains marginally important, X_1 (0.441) and X_2 (0.450) are now almost as important, approximately three times more important compared to X_3 (see Table 6.7).

6.5.2 Measuring the degree and impact of multicollinearity

In interpreting the regression variate, it is necessary to assess the degree of multicollinearity and to determine its impact on the results. The two most common measures for assessing collinearity are tolerance and variance inflation factor (VIF). Tolerance value is the amount of a variable unexplained by the other independent variables. Small tolerance values (and therefore large VIF values) denotes high collinearity. Generally accepted levels of multicollinearity are: tolerance values up to 0.10, corresponding to a VIF of 10. However, problems with multicollinearity may also be seen at much lower levels of collinearity and multicollinearity.

6.5.2.1 Diagnosing multicollinearity

In this study, tolerance values for the variables in the regression equation range from 0.633 (X_1) to 0.914 (X_3), indicating a rather narrow range of multicollinearity effects. (see Table 6.8). Likewise, the VIF values range from 1.580 (X_1) to 1.094 (X_3). Although none of these values indicate levels of multicollinearity that should seriously distort the regression variate, their effects on both the estimation and interpretation process need to be examined.

6.5.2.2 The effects of multicollinearity

(i) Multicollinearity creates "shared" variance between variables, thus decreasing the ability to predict the dependent measure as well as ascertain the relative contributions of each independent variable. As multicollinearity increases, the total variance explained decreases. Moreover, the amount of unique variance for the independent variable is reduced to levels that make estimation of their individual effects quite problematic particularly in explanation.

(ii) Multicollinearity can have substantive effect not only on the predictive ability of the regression models, but also on the estimation of the regression coefficients and their statistical significance tests. The impact of multicollinearity had been discussed during the estimation process (see Section 6.4.1).

(iii) Multicollinearity creates problems in interpretation. As multicollinearity occurs, the process for identifying the unique effects of independent variables becomes increasingly difficult. Since the regression coefficients represent the amount of unique variance explained by each independent variable, hence, as multicollinearity results in larger portions of shared variance and lower levels of unique variance, the effects of the individual independent variables become less distinguishable.

6.5.2.3 Multicollinearity in the regression model

Although bivariate correlations of > 0.70 may result in problems, even lower correlations may be problematic if they are higher than the correlations between the independent and dependent variables. In this study, a correlation of 0.575 between X_1 and X_2 represents "shared" variance of almost 33%. This can impact both explanation and estimation of the regression results.

Even values much lower than the suggested thresholds may result in interpretation or estimation problems, particularly when the relationships with the dependent measure is weaker. A correlation of 0.293 between X_1 and X_3 creates a "shared" variance of about 9%, almost as high as the explained variance of Y by X_3 (correlation = 0.365; variance explained ≈ 13%).

Table 6.8
Collinearity statistics of variables in the regression equation

Variable	Tolerance (>0.5)	VIF (<10)
X_1 (TCCt score)	0.633	1.580
X_2 (CTSR score)	0.699	1.494
X_3 (TCRt score)	0.914	1.094

With a bivariate correlation of 0.745 and 0.731 with the dependent variable respectively, both X_1 and X_2 are, by themselves, important predictors of representational competence, Y (see Table 6.2). However, with a bivariate correlation of 0.575, X_1 and X_2 are fairly highly correlated. Hence, when included in the regression equation, their part correlation are respectively X_1=0.351, X_2=0.386, and X_3=0.148. This has resulted in less unique explanatory power for each individual independent variable. For example, with a tolerance value of 0.633 for X_1, 36.7% of variance is explained by the other independent variables.

6.6 Validating the Regression Model

To ensure that the regression results are not specific only to the sample used in estimation but more generalizeble to the population, the regression model just derived needs to be validated. Obtaining another sample from the population and assess the correspondence of the results from the two samples is limited by such factors as time pressure, cost, or availability of respondents. Hence, in this study, the validity of the results shall be assessed in two different approaches. These are: (i) assessment of the adjusted R^2 and the degrees of freedom, and (ii) evaluating alternative regression models.

6.6.1 Assessment of the adjusted R^2 and degrees of freedom

From Table 6.6, the values of R^2 and adjusted R^2 are 0.714 and 0.709 respectively, a difference of only 0.005. Thus, it can be seen that with 3 predictor variables in the regression model, there is little loss in predictive power. This indicates a lack of over-fitting of the model.

In addition, in multiple regression, the degree of generalizability is represented by the degree of freedom (df), calculated as: df = sample size – no. of estimated parameters. The larger the df, the more generalizable are the results. With 3 variables in the model for a sample size of n=192, an adequate ratio of observations to variables (64:1) in the variate is maintained.

6.6.2 Evaluating other regression models

Estimation of the regression model using alternative methods such as the combinatorial approach or the confirmatory approach could help to validate the results obtained through stepwise multiple regression.

6.6.2.1 Estimating the regression model using the combinatorial approach

The combinatorial approach is primarily a generalized search process across all possible combinations of independent variables and the best-fitting set of variables is identified. With computerized estimation procedures, regression models can be generated almost instantly for any number of measures of predictive fit.

With 5 potential independent variables to be included, there are 5! or 120 possible combinations or regression models. However, at this stage, the researcher is only interested in regression models with different number of independent variables regardless of the arrangement. The independent variables were entered, one at a time, in decreasing order of their bivariate correlation with the dependent variable (see the correlation matrix in Table 6.2). Hence, TCCt score (X_1) was entered first, followed by CTSR score (X_2), TCRt score (X_3), LAQ score (X_4), and DSBT score (X_5). Using SPSS, five regression models were generated. Table 6.9 shows the multiple regression results using the combinatorial approach with different number of independent variables.

The regression model derived using the combinatorial approach also indicated that there were only three statistically significant predictors of representational competence. These are: TCCt score (X_1), CTSR score (X_2) and TCRt score (X_3). The 4th variable, LAQ score (X_4) is statistically not significant and contributes just 0.004 and 0.003 to the values of R^2 and adjusted R^2 respectively. The 5th variable, DBST score (X_5) is also statistically not significant, does not contribute to the R^2 value and instead, it causes the adjusted R^2 value to decrease while increasing the standard error of the estimate. Hence, in terms of prediction, model 3 which explains almost 71% of the variance of representational competence has the best overall model fit. See Table 6.9.

Table 6.9
Model summary of the combinatorial approach

Model Summary[f]

Model	Overall Model Fit				R^2 Change Statistics				
	R	R^2	Adjusted R^2	Std. Error of the Estimate	R^2 Change	*F* Change	df1	df2	Sig. *F* Change
1	.745[a]	.555	.552	5.643	.555	236.648	1	190	.000
2	.832[b]	.692	.689	4.704	.137	84.377	1	189	.000
3	.845[c]	.714	.709	4.547	.022	14.344	1	188	.000
4	.847[d]	.718	.712	4.526	.004	2.701	1	187	.102
5	.847[e]	.718	.711	4.537	.000	.146	1	186	.703

a Predictors: (Constant), TCCt
b Predictors: (Constant), TCCt, CTSR
c Predictors: (Constant), TCCt, CTSR, TCRt
d Predictors: (Constant), TCCt, CTSR, TCRt, LAQ
e Predictors: (Constant), TCCt, CTSR, TCRt, LAQ, DSBT
f Dependent Variable: TRCt

Table 6.10
Variables entered into the regression model
(stepwise regression and combinatorial approach)

Variable Entered	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics	
	B	Std. Error	Beta	*t*	Sig.	Zero-order	Partial	Part	Tolerance	VIF
Const.	-9.731	2.161		-4.502	.000					
X1	.933	.104	.441	8.988	.000	.745	.548	.351	.633	1.580
X2	.777	.082	.450	9.440	.000	.731	.567	.368	.669	1.494
X3	.421	.111	.155	3.787	.000	.365	.266	.148	.914	1.094

The variables entered into the regression equation using stepwise regression, combinatorial approach as well as confirmatory approach, all share the same sets of regression coefficients, correlations and collinearity statistics (see Tables 6.5, 6.10 and 6.12). Hence, in terms of explanation, there is no difference between the different regression models.

6.6.2.2 Estimating the regression model using the confirmatory specification

This approach includes all the variables at the same time (simultaneous regression). The potential impacts of multicollinearity on the selection of independent variables and the effects on overall model fit can be judged. This approach is particularly appropriate for validation purposes (Hair et al., 2006, p.228). Hence, it was used for validating the results of regression in this study. Tables 6.11 to 6.16 show the multiple regression results using the confirmatory approach with (i) all three, (ii) all four, and (iii) all five independent variables directly entered into the regression equation at one time.

Table 6.11
Model summary of the confirmatory approach (with 3 predictor variables)

Model Summary[b]

	Overall Model Fit				R^2 Change Statistics				
Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	R^2 Change	*F* Change	df1	df2	Sig. *F* Change
1	.845[a]	.714	.709	4.547	.714	156.405	3	188	.000[a]

a Predictors: (Constant), TCRt, CTSR, TCCt,
b Dependent Variable: TRCt

Table 6.12
Variables entered into the regression model - Coefficients[a]
(with 3 predictor variables)

Variable Entered	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics	
	B	Std. Error	Beta	*t*	Sig.	Zero-order	Partial	Part	Tolerance	VIF
(Const.)	-9.731	2.161		-4.502	.000					
X1	.933	.104	.441	8.988	.000	.745	.548	.351	.633	1.580
X2	.777	.082	.450	9.440	.000	.731	.567	.368	.669	1.494
X3	.421	.111	.155	3.787	.000	.365	.266	.148	.914	1.094

a Dependent variable: TRCt

Table 6.13
Model summary of the confirmatory approach (with 4 predictor variables)

Model Summary[b]

Model	Overall Model Fit				R^2 Change Statistics				
	R	R^2	Adjusted R^2	Std. Error of the Estimate	R^2 Change	*F* Change	df1	df2	Sig. *F* Change
1	.847[a]	.718	.712	4.526	.718	119.040	4	187	.000[a]

a Predictors: (Constant), LAQ, CTSR, TCRt, TCCt
b Dependent Variable: TRCt

Table 6.14
Variables entered into the regression model - Coefficients[a]
(with 4 predictor variables)

Variable Entered	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics	
	B	Std. Error	Beta	*t*	Sig.	Zero-order	Partial	Part	Tolerance	VIF
(Const.)	-14.064	3.403		-4.133	.000					
X1	.918	.104	.434	8.851	.000	.745	.543	.344	.628	1.593
X2	.776	.082	.449	9.466	.000	.731	.569	.368	.669	1.494
X3	.410	.111	.150	3.692	.000	.365	.261	.143	.910	1.099
X4	.079	.048	.065	1.644	.102	.178	.119	.064	.978	1.023

a Dependent variable: TRCt

Table 6.15
Model summary of the confirmatory approach (with 5 predictor variables)

Model Summary[b]

Model	Overall Model Fit				R^2 Change Statistics				
	R	R^2	Adjusted R^2	Std. Error of the Estimate	R^2 Change	*F* Change	df1	df2	Sig. *F* Change
1	.847[a]	.718	.711	4.537	.718	94.827	5	186	.000[a]

a Predictors: (Constant), DSBT, TCRt, LAQ, CTSR, TCCt
b Dependent Variable: TRCt

Table 6.16
Variables entered into the regression model - Coefficients[a]
(with 5 predictor variables)

Variable Entered	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics	
	B	Std. Error	Beta	*t*	Sig.	Zero-order	Partial	Part	Tolerance	VIF
(Const.)	-13.039	4. 339	.433	-3.005	.000					
X1	.916	.104	.452	8.792	.000	.745	.542	.342	.626	1.598
X2	.781	.083	.150	9.398	.000	.731	.567	.366	.655	1.527
X3	.410	.111	.062	3.688	.000	.365	.261	.144	.910	1.099
X4	.076	.048	-.015	1.562	.120	.178	.114	.061	.955	1.047
X5	-.119	.311	.433	-.382	.703	.036	-.028	-.015	.957	1.045

a Dependent variable: TRCt

Although with four and even five predictor variables (Tables 6.13 and 6.15), the overall relationship was still significant [F (4, 187) = 119.040, $p < 0.001$; F (5, 186) = 94.827, $p < 0.001$], only three of the predictor variables were statistically significant, at $p < 0.001$ (Tables 6.14 and 6.16). Hence, estimation of the regression model using the confirmatory specification (Table 6.12) reaffirms the regression results derived through both the stepwise procedure (see Table 6.5 - Variables entered into the regression model) and the combinatorial approach (see Table 6.10). That is: there were only three statistically significant predictors of representational competence which collectively explains about 71% of the variance of representational competence. These were: understanding of chemical concepts or prior knowledge I, (X_1), developmental level, (X_2) and understanding of chemical representations or prior knowledge II, (X_3).

6.6.3 Section summary

Unless estimated from the entire population, no regression model is the final and absolute model. This study is only an attempt to look for the best model. A model summary of the final regression model is given in Table 6.17.

Table 6.17
The final regression model - Model summary[d]

Predictor Variable	Multiple R	R^2	Adjusted R^2	R^2 change	b	β	Cumulative % of variance explained
X_1 TCCt score	.745[a]	.555	.552	.555	.933***	.441***	55.5
X_2 CTSR score	.832[b]	.692	.689	.137	.777***	.450***	69.2
X_3 TCRt score	.845[c]	.714	.709	.022	.421***	.155***	71.4

a Predictors: (Constant), TCCt
b Predictors: (Constant), TCCt, CTSR
c Predictors: (Constant), TCCt, CTSR, TCRt
d Dependent variable: TRCt

*** $p < .001$

The regression model with three independent variables (TCCt score, CTSR score, and TCRt score) explains more than 71% of the variance of representational competence. Prior knowledge (understanding of chemical concepts, X_1, and understanding of chemical representations, X_3) accounts for approximately 58% of the variance, while developmental level accounts for the remaining 14%. The regression model was a good fit (adjusted R^2 = 71%). The overall relationship was significant, F (3, 188) = 156.405, $p < 0.001$. With other variables held constant, TRCt score were positively related to TCCt score, CTSR score, and TCRt score. TRCt score increases by 0.933, 0.777, and 0.421 for every extra point of TCCt score, CTSR score, and TCRt score respectively. The effect of TCCt score, CTSR score, and TCRt score were all statistically significant at $p< 0.001$.

6.7 Linking the Regression Model with Theory

As mentioned earlier in Section 6.1, two objectives of the multiple regression analysis in this study were to predict representational competence and to identify the factor influencing representational competence. Since the predictor variables of representational competence had already been identified and prediction of representational competence discussed in Section 6.5.1.1, the main task of this section is to explain the influence of the three predictor variables on representational competence in terms of theory.

The final regression model in Table 6.17 shows that with only the 1st variable (X_1), almost 55.5% of the variance of representational competence is explained. The 2nd variable (X_2) explains almost 14% of the remaining variance, while the 3rd variable, X_3 only explains approximately 2% of the variance.

In terms of prediction, the regression coefficients (b) indicate that X_1 (b_1=0.933) is the most important predictor variable of representational competence while X_3 (b_3=0.421) is only marginally important.

However, the regression coefficients (β) allows for a direct comparison between coefficients as to their relative explanatory power of the independent variables. The beta coefficients (β) show that while X_3 remain marginally important, X_2 is just as important as X_1 and had substantial influence on representational competence, Y. Individually, bivariate correlations of X_1 and X_2 with Y are respectively r=0.745 and r=0.731, at $p<0.001$ (Table 6.2). Collectively, beta coefficients of X_1 and X_2 are respectively 0.441 and 0.450 (Table 6.17). Hence, it could be inferred that understanding of chemical concepts (X_1) is as important as developmental level (X_2) as factors influencing representational competence.

The theoretical and conceptual frameworks of the study proposed in Chapter 3 support the findings and the regression model that emerged from the findings. The theoretical framework repeatedly points to prior knowledge as the most important factor influencing representational competence, in particular prior knowledge I or understanding of chemical concepts (X_1). The theoretical framework also emphasized developmental level (X_2), as an important factor influencing representational competence, as well as understanding of chemical concepts.

6.7.1 Prior knowledge

With reference to the proposed theoretical framework in Chapter 3, the influence of prior knowledge on representational competence is important at several stages of information processing.

Firstly, within the sensory memory, prior knowledge helps to activate and control the perception filter (Figure 3.2a). Much of the sensory information will be filtered out if the learner does not possess the essential prior knowledge or concept due to missing schema in the LTM. For example: chemical representations such as chemical symbols and chemical formulae appear meaningless to a learner who does not understand basic chemical concepts like elements and compounds, atom and molecules. Likewise, chemical equations are just chunks of letters and numbers to those who do not understand what a chemical reaction is. Hence, prior knowledge within the LTM plays an important role in the selection process. In the context of this study, prior knowledge of the subjects includes understanding of basic chemical concepts or prior knowledge I and chemical representations or prior knowledge II. This is because representation of chemical concepts requires the learners not only to understand the chemical concepts and chemical representations involved, but also the

ability to link the three levels of representations, as well as to translate between representations at the same level and across the levels. These are aspects of representational competence assessed in this study.

Secondly, it is believed that linkages exist between WM and the LTM store. While processed information in the WM is passed to the LTM for storage, knowledge from prior learning is also being retrieved from the LTM to help with processing in the WM (Figure 3.2b). Problems can occur at this stage if there is a lack of prior knowledge due to missing schema in the LTM. According to Johnstone and Kellett (1980), items in the WM are handled as `chunks' of information, varying from single digits or characters to abstract concepts and complex formulae or structures. Integrating a large number of information bits into smaller number is one way of chunking. Schemata within the LTM enable learners to treat multiple elements as a single entity (chunking). The more a learner knows about a topic, the easier it is for chunking since chunking usually depends on some existing schemata in the LTM. See the example in Section 3.1.3.3 and Figure 3.2(c1). With a mean DSBT score of 7.33, the subjects of this study appear to possess large WM capacity, yet the regression model shows that WM capacity (X_5) is not a significant factor influencing representational competence while prior knowledge accounted for 58% of the variance of representational competence. Such findings suggest that low level of representational competence is most likely due to the lack of prior knowledge rather than a lack of WM capacity. Perhaps the type of information and the level of processing (Craik and Lockhart, 1972) also have affected information processing at this stage. Chunking might be easy for digits in the DBST but not for complex, unfamiliar chemical formulae, structures and equations. In addition, digits are easier to hold and manipulate, while making translations between chemical representations

could involve not only retrieving conceptual knowledge of chemical representations, but also creating mental images of them. There is a lot of information to hold and process, and processing occurs at a deeper level. Hence, the importance of prior knowledge overrides WM capacity as factor influencing representational competence.

Further more, according to Keig and Rubba (1993), making translation between representations is an information processing task that requires conceptual understanding about the representations. Finding of a positive, moderate correlation ($r=0.293$, $p<0.001$) between understanding of chemical concepts, X_1 (or prior knowledge I) and understanding of chemical representations, X_3 (or prior knowledge II) also supports the theory. The fact that both the variables (X_1 and X_3) are significant contributors to the regression model and that together they account for almost 58% of the variance of representational competence further supports the importance of prior knowledge in influencing representational competence.

6.7.2 Developmental level

Many concepts in chemistry are very abstract (Cantu & Herron, 1978). Herron (1978) also maintains that concepts such as atom and molecule should be considered formal in the Piagetian sense. Hence, it is likely that they cannot be totally understood without some formal reasoning. Findings of a bivariate correlation of $r=0.575$ at $p<0.001$ between understanding of chemical concepts (X_1) and developmental level (X_2) supports this claim. The fact that developmental level (X_2) is a significant contributor to the regression model which accounts for almost 14% of the variance of representational competence further supports the importance of developmental level in influencing representational competence.

Although chemical representations are apparently visual representations, they are also conceptual constructs and are therefore more abstract compared to pictorial diagrams. Chemical phenomena are interpreted at the microscopic level, requiring learners to think increasingly in abstraction. For learners who are unable to visualize and interpret molecular and symbolic representations, they only recognize the surface features of chemical representations and see or use representations as depictions. As a result, their understanding of chemistry tends to stay at the macroscopic or sensory level. Hence, formal operational thinkers who are capable of thinking abstractly not only can understand abstract chemical concepts more readily, but also can interpret chemical representations, logically use symbolic representations to represent and explain abstract chemical concepts, and can translate fluently between the three levels of representations, while those who remain in concrete operational stage may be limited in their understanding of chemical concepts, and in their representational competence.

6.7.3 Unexplained variance

The three predictor variables in the regression model explained almost 71% of the variance of representational competence. 29% of the variance remained unexplained. The percentage of variance explained could have exceeded 71%. This apparently lower percentage could be due to the following factors:

(i) Multicollinearity creates `shared' variance between variables. As multicollinearity increases, the total variance explained decreases as the amount of unique variance for each independent variable is reduced. In this study, a correlation of 0.575 ($p<0.001$) between X_1 and X_2 (see Table 6.2) shows X_1 and X_2 are moderately correlated, and represents `shared' variance of almost 33%. The

tolerance value of 0.633 for X_1 and 0.699 for X_2 respectively (Table 6.8) implies at least 30% of the variance of X_1 and X_2 is explained by the other independent variables. In addition, a correlation of 0.293 between X_1 and X3 creates a `shared' variance of about 9%, almost as high as the explained variance of Y by X_3 (r=0.365, variance explained =13%).

(ii) In selecting independent variables, specification error might have occurred, resulting in the omission of relevant variables influencing representational competence from the set of independent variables. This could bias regression estimates.

(iii) Exclusion of independent variables that are not statistically significant but might have practical significance could have caused the total variance explained to decrease. Although the partial correlation of both X_4 and X_5 are non significant at each stage of the regression analysis (Table 6.5), bivariate correlations of each variable shows X_4 is very weakly but positively correlated with Y (r=0.178, p<0.01), while the independent variable X_5 had non significant correlation (r=0.036) with the dependent variable. See Table 6.2. However, it is believed that the weak correlation between LAQ score, (X_4) and Y could be due to some meaningful learners switching to using a surface approach in certain learning tasks. This might be the cause of their apparently low LAQ score. Examination of the responses in the LAQ data of the 192 subjects show there were respectively 156, 123, and 110 students who chose either `A' or `B' as their responses for items No. 3, 5 and 19. These responses relate to memorisation, a phenomenon associated more with a surface approach (see *Appendix 19*). Hence, in the LAQ, such items were placed under the rote subscale (Table 4.7) and the score is reversed (Table 4.8), leading to lowest score for response `A'. A check on the reliability of the LAQ score for this study reveals an alpha coefficient of

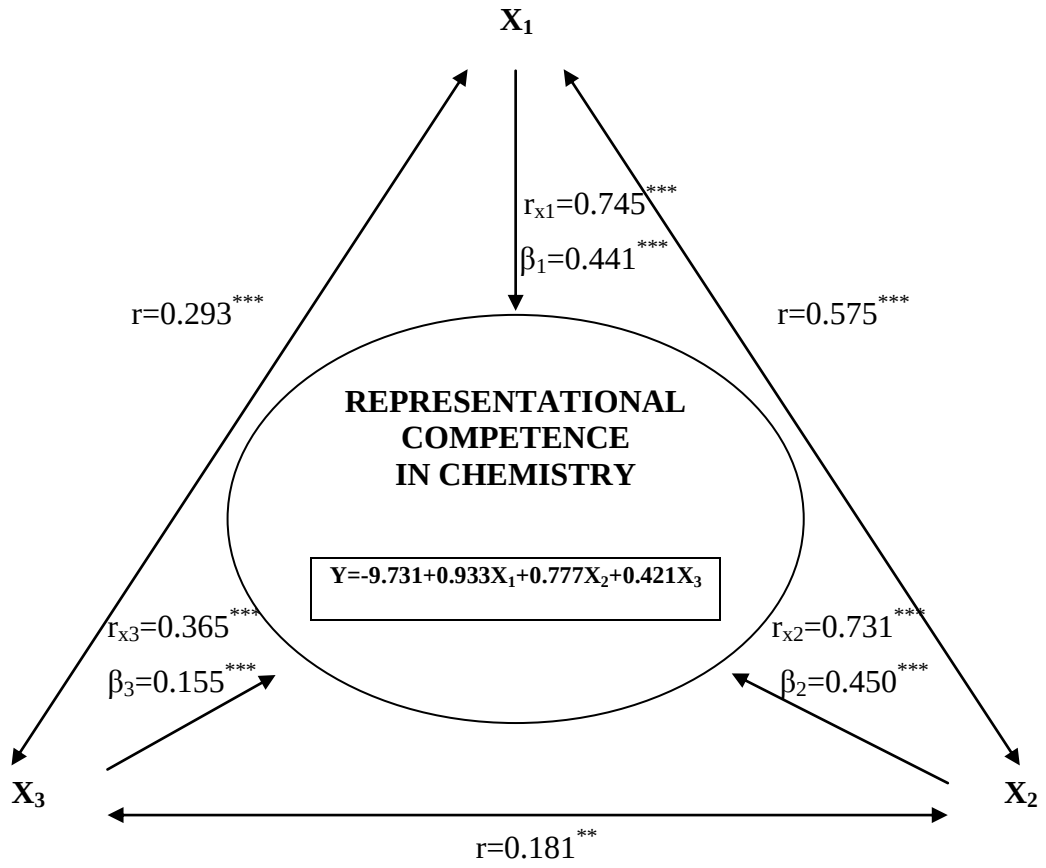
0.77 for the meaningful learning subscale while that of the rote learning subscale is relatively much lower at alpha=0.47 (*Appendix 19a*). This reflects internal inconsistency of items within the rote learning subscale.

The weak correlation between working memory capacity, X_5 and Y is probably caused by the almost uniform distribution and generally high DSBT scores (mean=7.33, s.d.=1.079). Further research is needed.

6.8 Chapter Summary

The model that emerged from the findings of this study shown in Figure 6.6 summarizes the variables influencing representational competence and the inter-relationships among them.

In Chapter 7, the implications and conclusion of the study will be put forward.



Independent variables

X_1 = Understanding of chemical concepts (prior knowledge I)

X_2 = Developmental level

X_3 = Understanding of chemical representations (prior knowledge II)

Dependent variable, Y = Representational competence

b = regression coefficient (In the equation: b_0=-9.731, b_1=0.933, b_2=0.777, b_3=0.421)

β = standardized regression coefficient

r = bivariate correlational coefficient

*** p<0.001; ** p<0.01

Figure 6.6: The emerging model