Leadership Effects on Teaching

experience. As for the female gender, 80 out of the 114 female respondents had less than five years of work experience, followed by no work experience (16 or 14.0%). 23 (20.2%), 14 (12.3%), and the rest of the nine respondents (7.9%) had 5-9 years, 10-14 years, more than 15 years of work experience.

Last, in terms of location of residence, a large majority of respondents were from Hong Kong, accounting for 125 out of the 162 respondents or 77.2%. It was followed by those living in Mainland China (35 or 21.6% of all respondents). There were only two overseas respondents (1.2%). As for the distribution of the male gender, almost all (89.6%) or 43 of them were from Hong Kong. The rest of them (5 or 10.4%) were from Mainland China. As for the female gender, 82 or 71.9% of them were from Hong Kong, followed by 30 or 26.3% of them from Mainland China. The rest of the small minority were from overseas countries.

It is worth noting that the gender imbalance of respondents is one of the limitations of this study (to be discussed further in Chapter 6 Conclusion and Recommendation). Two reasons may explain it. The first one is inherently more female master's degree students in the School of Accounting and Finance (SAF) in PolyU. The gender proportion is about 40 (Male): 60 (Female) for master's degree students in SAF in PolyU. However, the first reason cannot fully explain such a polarizing gender proportion in respondents. For instance, the second reason may be that all respondents were recruited through e-mail. They then filled in the questionnaire through an online form. Students from the master's program can choose to respond to the invitation or not voluntarily. Therefore, there was no way to control the demographic characteristics of the respondents.

4.2. Mean and Standard Deviation, One-way ANOVA Analysis

There are two more analyses: mean and standard deviation analysis and one-way ANOVA analysis. The result of the mean and standard deviation analysis is shown in Table 4-2.

Table 4-2 Mean and Standard Deviation Analysis

	Mean	SD	Skewness	Mean score per item	Average SD per item
TIL	113.99	15.91	-.90	5.70	.80
IM	16.85	3.08	-.38	5.62	1.03
EM	16.99	3.29	-.98	5.66	1.10
PAP	16.17	3.04	-.85	5.39	1.01
AL	15.32	3.07	-.11	5.11	1.02
IC	27.91	5.33	-1.35	5.58	1.07
CL	10.23	2.04	-.25	5.12	1.02

Source: Survey's data analysis

Note: SD---Standard Deviation, TIL--- TIL, IM--- IM, EM--- EM, Instructor's Credibility--- IC, Perceived Academic Performance--- PAP, Cognitive Learning--- CL, Affective Learning--- AL.

Table 4-2 shows the mean score, standard deviation, and skewness of the seven constructs in this study. The scores of all constructs are negatively skewed, meaning that the mean is smaller than the median and mode. Overall, the mean score for each measurable item for each construct is high on a seven-point Likert Scale. TIL, IM, EM, and IC recorded an average score of about 5.60. Other constructs: PAP, AL, and CL had high average scores of 5.39, 5.11, and 5.12. The average standard deviation for each item was low for all constructs, ranging between .80 and 1.10. It means that most of the scores of about 68% are within the range of 4 to 6.

The result of the one-way ANOVA analysis is shown in Table 4-3. None of the one-way ANOVA is statistically significant, indicating **no significant differences** in scores in different demographic classifications.

Table 4-3 One-way ANOVA analysis

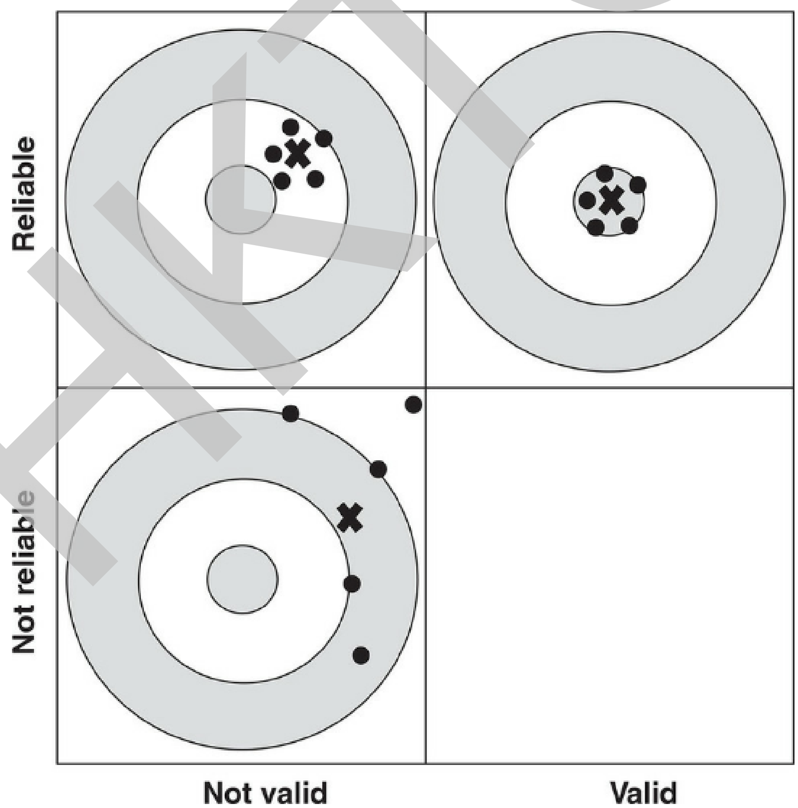
	Gender	Age	Work Experience	Location of Residence
TIL	.241	.434	.499	.739
IM	.055	.239	.328	.910
EM	.297	.771	.496	.755
PAP	.427	.624	.303	.575
AL	.248	.274	.395	.542
IC	.125	.672	.635	.621
CL	.641	.631	.425	.710

Source: Survey's data analysis

4.3. Validity and Reliability of Instruments for Measuring Constructs

It is crucial to discuss the concept of validity and reliability after discussing the demographic characteristics. In statistics and research studies, the validity of instruments refers to whether a research instrument measures what it is supposed to measure (i.e., a research construct) (D. Cooper & P. S. Schindler, 2014). Therefore, validity means whether various indicators in a questionnaire or other instruments measure a construct (Hair et al., 2017). In other words, in a questionnaire, there are many indicators for a construct. Therefore, an indicator must be representing the nature of a construct. Reliability of instruments refers to how reliable a research instrument is in measuring the research construct (D. Cooper & P. S. Schindler, 2014). Therefore, reliability means whether the various indicators accurately measure a construct (Hair et al., 2017). Graphical representation as extracted in Hair et al. (2017) is shown in Figure 4-2.

Figure 4-2 Validity and Reliability in Graphical Representation



Adopted from Hair et al. (2017).

The below measures of various types of validity and reliability were applied to assess the quality of the measurement models.

4.3.1. Validity of Instruments

The major type of validity measure for instruments **construct validity** (i.e., *convergent validity* and *discriminant validity*), **internal validity**, and **external validity** (D. Cooper & P. S. Schindler, 2014). **Construct validity** refers to how the indicators measure what it is purported to measure (D. Cooper & P. S. Schindler, 2014). There are two major types of construct validity: *convergent validity* and *discriminant validity*. *Convergent validity* refers to how the scores on one scale correlate with the scores on another scale designed to measure the same construct (Bergin, 2018). In other words, it means that when there are two or more different scales measuring the same constructs, it is crucial to know whether one scale correlates with another measurement scale. If the correlation between the one scale and another scale is high, the scale has *convergent validity*. *Discriminant validity* refers to the degree to which scores on a scale do not correlate with scores from scales designed to measure different constructs (D. Cooper & P. S. Schindler, 2014). In other words, there are many sub-scales in measuring different constructs in a questionnaire instrument, it is crucial whether measurement of a construct such as "perceived academic performance" is not correlated with the measurement of another construct such as "affective learning." If the correlation between the two scales is not high, the two scales have discriminant validity.

In terms of measuring convergent validity, an ideal way is to calculate the correlation of a proposed test for a construct with an established one (D. Cooper & P. S. Schindler, 2014). However, it is not always possible since an instrument for a construct may be the first of a kind. It means that there may not be a competing instrument for measuring the correlation of the instrument with a competing one. Alternatively, PLS-SEM applies *outer loadings* and the *average variance extracted (AVE)* to determine the convergent validity (D. Cooper & P. S. Schindler, 2014).

First, the *outer loading* analysis uses the logic that a measurable construct item is an alternative approach to measure the same construct. The measurable item has convergent validity if the measurable item covers or shares a high proportion of variance of the construct. In other words, outer loading is measured by dividing the amount of variance captured by an indicator by the amount of total variance of a construct (Hair et al., 2017). In other words, outer loading represents the amount of variation of an item explained by the construct (i.e., measured by the total variance of the construct). An acceptable outer loading indicates sufficient communality of a measurable item. An outer loading of above .70 is considered acceptable (Hair et al., 2017).

The second method to ensure convergent validity is *AVE*. *AVE* represents the grand mean value of the square loadings of the indicators related to a construct (i.e., the sum of the squared loading divided by the number of indicators). An AVE value of .50 or higher indicates that the construct, on average, explains more than half of the variance of its indicators. Using the same logic that an indicator is an alternative construct measure, a high average value of indicator variance signifies that the indicators explain the construct.

Discriminant validity is usually measured by conducting a simple correlation coefficient analysis between the various constructs in a research model or the constructs that are supposed to be discriminated (D. Cooper & P. S. Schindler, 2014). In PLS-SEM, researchers use a more sophisticated method called heterotrait-monotrait ratio (HTMT) to determine discriminant validity. In short, HTMT is the ratio of the between-trait correlations to the within-trait correlations. Specifically, HTMT is the mean of all correlations of indicators across constructs measuring different constructs relative to the geometric mean of the average correlations of indicators measuring the same construct (Hair et al., 2017). It is a more advanced method of determining the discriminant validity since the traditional correlation coefficient method relies on the summation of all indicator scores of a sample between the two constructs to arrive at the correlation figures. It is more rudimentary than HTMT. HTMT is more sophisticated by considering the construct level and the indicator-level correlations across constructs and within a construct. As a rule of thumb, an HTMT of above .85 or .90 suggests a lack of discriminant validity (Henseler et al., 2015).

On the other hand, the *internal validity* and *external validity*. *Internal validity* is the degree to which a study establishes a concrete causal relationship between the independent variable (IV) and dependent variables. It depends on the study procedures and how rigorously it is performed (D. R. Cooper & P. S. Schindler, 2014). *External validity* is the degree of how well the study outcome can be expected to apply to other settings (D. Cooper & P. S. Schindler, 2014). The positive influence of the *Internal validity* of this study is that the study has a high level of R Square between TIL to the two mediators and four outcomes, ranging from .323 to .558 (See Table 4-8). It represents that research IV has certain explanatory power on the mediators and the outcomes. There may not be too many confounding variables influencing the outcomes. However, the negative influence on the *Internal validity* of this study is that the selection of samples. First, the selection was not random. It only involved distributing questionnaires to postgraduate students of PolyU AF for invitation. It did not randomly select all postgraduate students to fill in and invite them to reply. It is one of the limitations of this study.

On the other hand, external validity may not be highly generalizable to all postgraduate students in Hong Kong. It did not acquire the contact information of all postgraduate students in Hong Kong and randomly selected and invited them to participate in the study. Instead, the population was only the postgraduate students of PolyU AF. Overall, it did not extend to the entire population of postgraduate students studying in universities situated in Hong Kong. It thus limits the generalizability of this study to entire Hong Kong. It is another limitation of this study. Therefore, the evaluation of internal and external validity reveals two limitations for this study. These limitations limit the generalizability of this study.

4.3.2. Reliability of instruments

Reliability is a measurement of how accurate a scale for measuring a construct is. The most common measure is internal consistency reliability using Cronbach's Alpha. Cronbach's Alpha measures how closely related a set of items are as a group (Ott & Longnecker, 2015). Cronbach's Alpha is acceptable at .70, good at .80 and excellent at .90. Cronbach's Alpha **internal consistency reliability** measure and **composite reliability (CR)** are used to measure reliability in PLS-SEM.

Composite reliability considers the different outer loadings of the indicators. It is calculated by summing the outer loadings of indicators and dividing them by the sum of itself and the variance of the measurement errors (Hair et al., 2017). It is a better measure than Cronbach's Alpha since it is not sensitive to the number of items in the scale. CR of .60 to .70 is acceptable in exploratory research. For more advanced stages of research, the value between .70 and .90 are considered satisfactory. The above .90 is not desirable because they indicate that all indicators are measuring the same thing, rendering it meaningless (Hair et al., 2017).

4.3.3. Multicollinearity and Model Fit

Multicollinearity is a statistical phenomenon in which one predictor variable in a path model can be predicted linearly from the others with a substantial degree of accuracy. It indicates that, even though the criteria for validity and reliability are satisfied, one predictor can still be predicted linearly by another predictor. It indicates that the quality of the data collected may be problematic. Multicollinearity is measured by a variance inflation factor (VIF) (Hair et al., 2017). The formula of VIF is shown below:

$$VIF = \frac{1}{(1 - R_J^2)}$$

$Where$: R_J^2=the coefficient of determination of regression of explanator or construct j on all other explanators.

As a rule of thumb, multicollinearity of ≥10 indicates a serious multicollinearity problem (Hair et al., 2010).

Model Fit is a measure of the discrepancy between the observed values and the expected values. It is crucial since it indicates whether the hypotheses go too far from the observed values. In its simplest way, model fit is measured by Pearson's chi-square test, which summates observed values minus expected values and divides them by the expected value. In PLS-SEM, there are two measures of model fit: *standardized root means square residual (SRMR)* and *root mean square residual covariance (*RMS_{theta}*)* (Hair et al., 2017). SRMR measures the root mean square discrepancy between the observed correlations and the model-implied correlations

(Hair et al., 2017). SRMR value below .08 indicates a good fit for CB-SEM using IBM SPSS AMOS, but no threshold value is introduced in a PLS-SEM context yet (Hair et al., 2017). RMS_{theta} measures the root mean square discrepancy between the observed covariances and the model-implied covariance (Hair et al., 2017). RMS_{theta} of below .12 is a conservative threshold for a good fit, while a higher value indicates a comparative lack of fit (Henseler et al., 2014).

4.4. Evaluation of Measurement Models

This section presents the evaluation of construct validity, discriminant validity, reliability, multicollinearity, and model fit. They are conducted for the evaluation of measurement models.

There are two tests for convergent validity: outer loadings and AVE. The test for discriminant validity is HTMT. There are two more tests for reliability: Cronbach's Alpha Internal Consistency Test and Composite Reliability. These analyses are shown in Table 4-4, Table 4-5, Table 4-6, and Table 4-7.

Table 4-4 Results of Outer Loadings

	AL	CL	EM	IC	IM	PAP	TIL
AL1	0.884						
AL2	0.930						
AL3	0.935						
CL1		0.946					
CL3		0.829					
EM1			0.827				
EM2			0.867				
EM3			0.865				
IC1				0.923			
IC3				0.897			
IC4				0.902			
IC6				0.891			
IC7				0.853			
IM1					0.878		
IM2					0.881		
IM3					0.904		
PAP1						0.770	
PAP2						0.945	
PAP3						0.903	

TIL1	0.757
TIL2	0.772
TIL3	**0.689**
TIL4	**0.694**
TIL5	0.834
TIL6	0.780
TIL7	0.748
TIL8	0.843
TIL9	0.786
TIL10	**0.650**
TIL11	0.763
TIL12	0.776
TIL13	0.705
TIL14	0.774
TIL15	0.810
TIL16	0.833
TIL17	0.880
TIL18	0.882
TIL19	0.883
TIL20	0.718

Source: Survey's data and analysis
Note: TIL--- TIL, IM--- IM, EM--- EM, Instructor's Credibility--- IC, Perceived Academic Performance--- PAP, Cognitive Learning--- CL, Affective Learning--- AL

As for outer loading in Table 4-4, except for TIL3, TIL4, and TIL10, the outer loadings for all indicators are over .70, which meet the minimum requirements for outer loadings. Besides, there are originally 42 indicators. However, in the final analysis, IC2, IC5, and CL2 are deleted since their outer loadings are substantially lower than .70. According to the guidance of Hair et al. (2017), researchers should consider excluding those indicators with outer loadings of lower than .70. If the outer loadings of the other indicators in the same construct become higher in another round of analysis, researchers have to confirm the exclusion. If it is not, they should put those indicators back to the analysis. In the current research, the outer loadings for the other indicators in the construct of IC and CL become higher after excluding IC2, IC5, and CL2. Therefore, the researcher then excludes IC2, IC5, and CL2 permanently from the analysis. In contrast, TIL3, TIL4, and TIL10 are not excluded in the final analysis since excluding them does not make the outer loadings of other indicators higher.

Table 4-5 Results of AVE, Cronbach's Alpha, and Composite Reliability

Construct	AVE	Cronbach's Alpha	Composite Reliability
TIL	.611	.966	.969
IM	.789	.867	.918
EM	.728	.857	.889
IC	.798	.943	.952
PAP	.768	.875	.908
CL	.791	.912	.883
AL	.841	.933	.941

Source: Survey's data and analysis
Note: TIL--- TIL, IM--- IM, EM--- EM, Instructor's Credibility--- IC, Perceived Academic Performance--- PAP, Cognitive Learning--- CL, Affective Learning--- AL

As for the AVE results, all constructs had AVE of higher than .50, satisfying the criteria for convergent validity. On the other hand, the Cronbach's Alpha results for internal consistency reliability are satisfying since all are higher than .70. However, for composite reliability, TIL, IM, IC, PAP, and AL have figures higher than .90. Thus, according to the guidance of J. F. J. Hair et al. (2017), it is not a desirable result since the indicators may not just measure the same thing to the respondents. However, though it is one of the limitations of this study, it is just a minor limitation since all values indicate at least an acceptable internal consistency reliability. In other words, there is no sign of unacceptable internal consistency reliability.

Table 4-6 Results of HTMT

	TIL	IM	EM	IC	PAP	CL	AL
TIL	-						
IM	.780	-					
EM	.620	.811	-				
IC	.654	.517	.489	-			
PAP	.696	.847	.676	.567	-		
CL	.596	.689	.485	.371	.921	-	
AL	.521	.600	.517	.291	.741	.855	-

Source: Survey's data and analysis
Note: TIL--- TIL, IM--- IM, EM--- EM, Instructor's Credibility--- IC, Perceived Academic Performance--- PAP, Cognitive Learning--- CL, Affective Learning--- AL

As for discriminant validity in Table 4-6, apart from CL>PAP, none of the construct pairs show an HTMT of higher than .85 or .90, demonstrating adequate discriminant validity. The close relationship between CL and PAP is understandable since a higher level of cognitive learning means higher information processing skills. It is thus conducive to a better perception of academic performance.

Table 4-7 Multicollinearity Analysis by VIF

Indicators	VIF
AL1	2.662
AL2	3.082
AL3	3.285
CL1	1.569
CL3	1.569
EM1	2.058
EM2	2.272
EM3	1.563
IC1	3.956
IC2	**5.351**
IC3	4.479
IC4	4.331
IC5	3.078
IM1	1.952
IM2	2.457
IM3	2.666
PAP1	1.594
PAP2	3.867
PAP3	3.231
TIL1	4.429
TIL2	3.814
TIL3	2.927
TIL4	2.760
TIL5	**5.168**
TIL6	**5.035**
TIL7	3.361
TIL8	4.971
TIL9	3.500
TIL10	2.696
TIL11	3.655
TIL12	3.843
TIL13	3.174
TIL14	**6.147**
TIL15	**6.564**
TIL16	**8.681**
TIL17	**6.616**
TIL18	**6.462**
TIL19	**6.070**
TIL20	3.104

Source: Survey's data and analysis Note: TIL--- TIL, IM--- IM, EM--- EM, Instructor's Credibility--- IC, Perceived Academic Performance--- PAP, Cognitive Learning--- CL, Affective Learning--- AL

As for multicollinearity analysis in Table 4-7, all indicators have VIF scores of lower than 10, indicating that the measurement model does not have serious multicollinearity problems. Some indicators such as IC2, TIL5, TIL6, TIL14- TIL 19 have VIF scores of more than 5, indicating a higher level of multicollinearity. However, they are still acceptable since they are not higher than ten, as indicated by (Hair et al., 2010). Last, the model fit scores of SRMR and RMS_{theta} are .126 and .188, indicating a less desirable model fit since the SRMR and RMS_{theta} are higher than .08 and .12. It is one of the limitations of this study.

Overall, despite higher composite reliability scores for some constructs and higher multicollinearity scores for some indicators, the measurement model is still valid and reliable since these problems are not serious. The problem is not serious since convergent and the measures of AVE and HTMT indicate discriminant validity. Second, despite some Cronbach's Alpha values and Composite reliability measures of higher than .90, it only indicates that some indicators in constructs may be measuring the same thing. It does not defeat that the internal consistency reliability is acceptable and indicates reliability. Third, the SRMR and RMS_{theta} are higher than the threshold. They do not greatly exceed the threshold. Therefore, despite some minor problems in the internal consistency reliability and model fit, it does not indicate that the model is not valid and reliable.

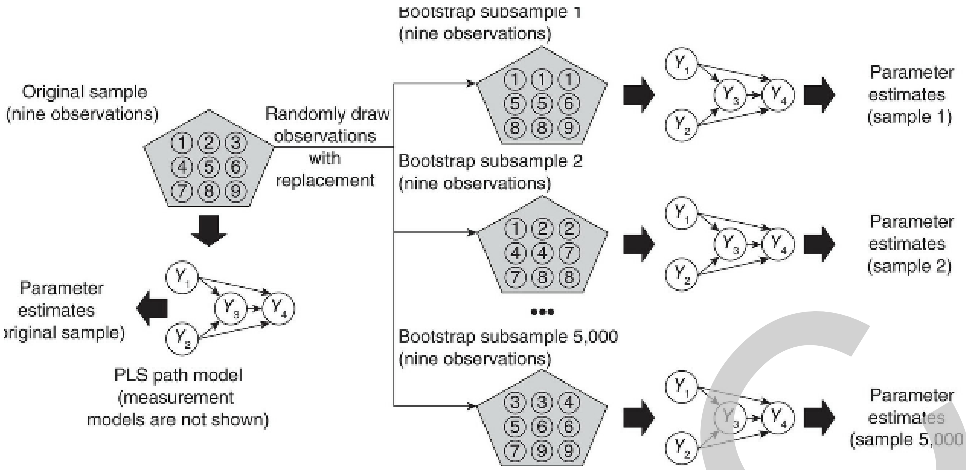
4.5. The Path Model Presentation

The results of the path model analysis are shown in Figure 4-3. In addition, the path model's total effects and specific indirect effects are shown in Table 4-8 and Table 4-9. First, the path model is obtained by the path analysis in PLS-SEM. It is about the strength of the relationship between IV, mediators, and the outcomes. It uses bootstrapping to indicate whether these relationships are statistically significant (Hair et al., 2017).

statistically significantly related to AL, CL, and PAP at .280, .423, and .496. EM is statistically significantly related to IC at .133. Therefore, as reconciled in Table 4-9, the significant indirect effect from TIL to AL, CL, and PAP through IM is .204, .307, and .361. The significant indirect effect from TIL to IC through EM is .076. Table 4-10 summarizes the direct and indirect effect between TIL to the four outcomes.

The following paragraphs explain the use of P-value, R-Square, and Q-Square. **P-value** is for determining the significance level of the result. A general rule is that a statistical result is significant at a 5% significance level or $p<.05$. $p<.10$ represents that there is a 10% of probability that the statistical result is not significant. $P<.05$ and $p<.01$, therefore, represent 5% and 1% probability. 10% significance level is often considered too lax since it means a 10% probability of having type 1 (i.e., rejecting a true null hypothesis) errors. However, a 1% significant level is considered too stringent in rejecting a null hypothesis. Therefore, a 5% significance level is generally used (Ott & Longnecker, 2015). A significance level of PLS-SEM is derived by a process called bootstrapping. Bootstrapping is a resampling approach that draws random samples (with replacement) from the data. Replacement means that each time an observation is drawn randomly from the sampling population, it is returned to the sampling population before the next observation is drawn. The essence of bootstrapping is to use these samples to estimate the path model multiple times under a slightly changed data constellation randomly (Hair et al., 2017). It then estimates the p-value for significance testing. It is required since PLS-SEM is an exploratory statistical analysis technique that is more suitable to test unknown data patterns instead of more established data patterns. Using only non-random and established data patterns will result in errors in estimating the p-value and t-value. Therefore, it is assumed to be a distribution-free multivariate data analysis technique without relying on distributional assumptions (Hair et al., 2017). A graphical representation of bootstrapping is as follows in Figure 4-4.

Figure 4-4 Graphical Representation of Bootstrapping



Source: Hair et al. (2017)

Therefore, to estimate the t-value or p-value, bootstrapping is a technique that is needed to generate as many random samples as possible. In other words, making the data pattern more random is necessary to estimate the t or p-value.

In this study, as recommended by Hair et al. (2017), 5,000 bootstrap samples, each of which includes the same number of cases as the number of observations in the original data set, are drawn to arrive at the significance level. Table 4-10 shows that TIL is statistically significant to all two mediators and the four outcomes. On the other hand, one of the mediators---IM is statistically significant to three outcomes: AL, CL, and PAP. However, another mediator-EM is only statistically significant to IC. The path effect of EM to IC is also much smaller than that of IM. Judging only by the p-value estimation, IM seems to be a much stronger mediator than EM.

R Square is also known as the coefficient of determination. It determines how well a construct is explained by another construct (Ott & Longnecker, 2015). It is calculated by the square of the correlation coefficient or r. Researchers often use the adjusted R Square as the final R Square calculation(Ott & Longnecker, 2015). A high R Square means more of the variances of a construct is predicted by another construct. TIL has the highest R Square to PAP and IM in this study, followed by IC, CL, EM, and AL. All R Square measurements are statistically significant.

An associated measurement of R Square is **F Square or effect size**. Effect size determines the level of influence on another construct once the former construct is removed from the calculation. For example, the following formula calculates it:

$$F\ Square = \frac{(R\ Square\ included - \text{R Square excluded})}{(1 - R\ Square\ included)}$$

In the above formula, R Square included is the R Square that encompasses the influence of a construct on another construct. In contrast, R Square excluded is the R Square that excludes the influence of the subject construct to another construct.

A higher level of F Square indicates a higher level of predictive accuracy. According to Cohen (1988), .02, .15, and .35 indicate small, medium, and large effect sizes. Per Table 4-8, the effect size of TIL->IM is the largest at 1.112, followed by EM at .491. It means that IM and EM have large effects on the influence of TIL on other constructs. It is normal since TIL is the only independent variable and both IM and EM are the mediators. In other words, only through IM and EM can TIL influence the other four constructs or outcomes.

As for the rest of them, the effect sizes of IM to CL and PAP are moderately large and large, respectively. Besides, the significant effect size of IM to AL is slightly lower than the medium threshold suggested by Cohen (1988). On the other hand, none of the F Square from EM is statistically significant, indicating that EM does not influence the change of R Square or coefficient of determination of any of the four outcomes. In other words, the effect size of IM is dramatically larger than that of EM. Therefore, it contributes much more to the four outcomes.

Last, **Q Square**, different from R Square, is a measurement of predictive relevance. It indicates whether an independent variable is relevant or, in other words, appropriate in predicting the outcomes. For example, table 4-8 shows that TIL has the greatest influence on PAP and IM, followed by IC, CL, AL, and EM. In other words, TIL is the most relevant in predicting IM and PAP. Apart from Q Square, another measure called q square indicates the effect size of a construct in influencing the Q Square of another construct. Similar to F Square, q square is calculated by the following formula:

$$q\ square = \frac{(Q\ Square\ included - \text{Q Square excluded})}{(1 - Q\ Square\ included)}$$