Volume: 3 Issues: 8 [June, 2018] pp.57-66]
Journal of Information System and Technology Management
eISSN: 0128-1666
Journal website: www.jistm.com

DATA CLEANING FOR THE EVALUATION OF VIRTUAL LEARNING ENVIRONMENT SUCCESS AMONG TEACHERS

Hapini Awang[1]
Zahurin Mat Aji[1]
Wan Rozaini Sheik Osman[1]

[1]School of Computing, College of Arts & Sciences, Universiti Utara Malaysia, Malaysia.

Accepted date: 29 January 2018 **Published date**: 2 July 2018

To cite this document:

Awang, H., Aji, Z. M., & Osman, W. R. S. (2018). Data Cleaning for The Evaluation of Virtual Learning Environment Success Among Teachers. *Journal of Information System and Technology Management*, *3*(8), 57–66.

Abstract: *This article explains the data cleaning procedures for the evaluation of Virtual Learning Environment (VLE) success among Malaysian teachers. Data cleaning is essential step in any research to ensure that the produced data is usable, valid and reliable for testing the research framework. The data collection of this study was done among the teachers across four states of Perlis, Kedah, Penang and Perak. During the data cleaning procedures, seven main issues have been considered which include missing data, outliers, linearity, normality, homoscedasticity, multicollinearity and common method variance. As a result, a dataset consists of 643 good attributes' cases were produced. All the assumptions for multivariate analysis had been met. Besides, this dataset has been proved to be robust and ready for further statistical examinations.*

Keywords: *Data Cleaning, Data Preparation, Data Screening, E-Learning, Information System Success, Virtual Learning Environments*

Introduction

Virtual Learning Environment (VLE) is a type of e-learning system that is implemented in various educational institutions such as universities, training centers and schools to systematically support the online learning and administrations (Mueller & Strohmeier, 2011). The reputation of VLE as a well-established technology has positively facilitated the process of knowledge retrieval and online learning management (Nor Fadzleen et al., 2013). Furthermore, the VLE technology has significantly shifted the nature of traditional learning in six aspects; time, place, space, technology, interaction, and control (Piccoli, Ahmad, & Ives, 2001). In Malaysia, the implementation of VLE platform, known as Frog VLE is a part of educational information systems' provision to support the digital teaching and learning, as well as education management. However, despite various benefits offered by the Frog VLE, a recent evidence has demonstrated that some teachers in Malaysia refuse to continue using the system, although they are agreed on the benefits offered by the system (Cheok & Wong, 2016).

Consequently, this issue has been identified as the main contributor for the Frog VLE low usage statistics and signifies its vulnerability to system's failure (Johari & Siti Norazlina, 2010; Kementerian Kewangan Malaysia, 2014; Nor Azlah et al., 2014). Therefore, the study to evaluate VLE success among teachers in Malaysia has been initiated, which the data collection was done starting from July to October 2017.

For the main analysis, the Partial Least Squares-Structural Equation Modeling (PLS-SEM) will be used. Nonetheless, to ensure that the study will produce a valid and accurate result, several pre-requisite procedures need to be taken before the data analysis. Among the vital procedures is data cleaning. Data cleaning is an important step before embarking on any multivariate statistical analysis. Raw data, especially the primary data that are collected from respondents will mostly contain flaws or contaminations, which could be resulted from several factors, including respondents, instrumentation and sampling errors. In this sense, the proper considerations of several issues are essential, especially to establish an honest analysis of the data (Tabachnick & Fidell, 2007). Therefore, this paper presents the analysis of data cleaning for the investigation of VLE success amongst teachers.

Research Framework

Upon the literature analysis on the related Information Systems (IS) theories and models, it is found that the VLE success could be examined using the updated DeLone and McLean IS Success Model (D&M) (DeLone & McLean, 2003). Previous studies proved that this model fits all the measurement for IS success evaluation (Al-Debei, Jalal, & Al-Lozi, 2013; Mohammadi, 2015; Zhu, Lee, Kuo, & Lin, 2013). DeLone and McLean (2003) suggest that the application on D&M should consider the context of IS under investigation and refine the model to suit the variability of IS research disciplines. Therefore, the current study develops a research framework based on the updated D&M, with a number of modifications including the inclusion of Workload (WL) as a moderator and Intention to Use (ITU) as the predictor for Use (U). A part from that, the research framework consists of existing constructs of D&M; Information Quality (IQ), System Quality (SyQ), and Service Quality (SeQ), User Satisfaction (US) and Net Benefits (NB), as shown in Figure 1.

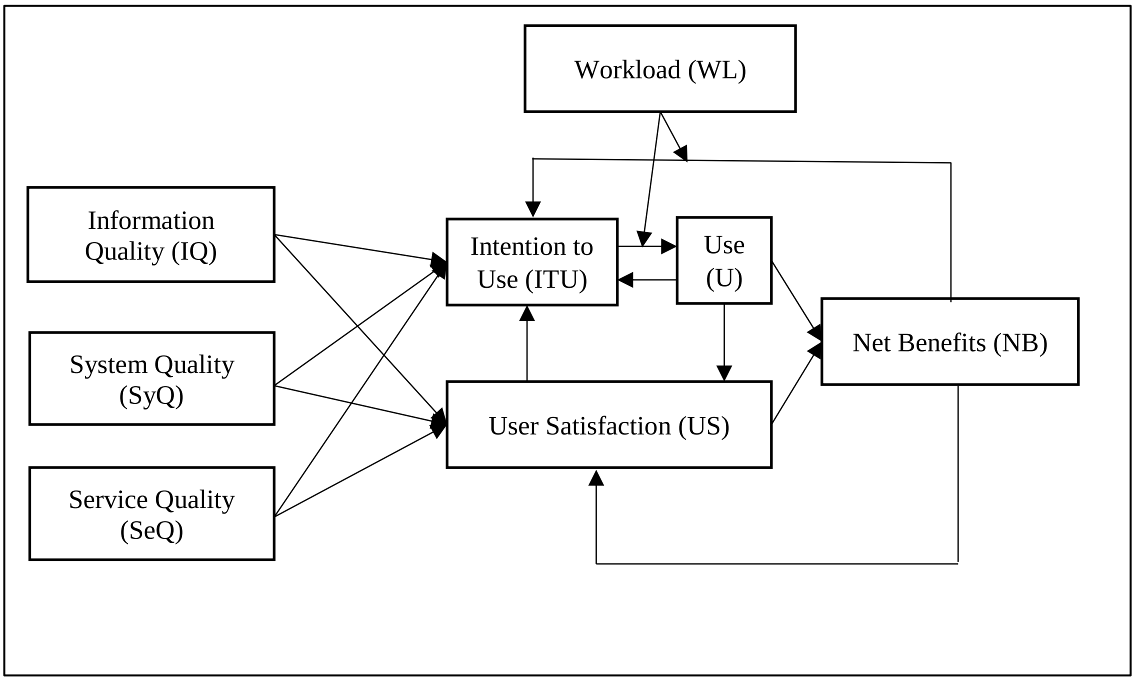


Figure 1: Reseach Framework

Questionnaire Design

Consists of 52 items, the questionnaire of this study was designed to measure eight constructs; IQ – seven items, SyQ – seven items, SeQ – nine items, ITU – four items, U – nine items, US – four items, WL – seven items, and NB – five items. All the items of this questionnaire were adopted from various sources, which have gone through the validation process by seven experts in the study field. Before conducting the pilot study, the items in the questionnaire were revised. To ensure that this questionnaire reaches the target respondents, one filtering question is included in the demographic section to identify the non-users of VLE. For every construct, the measurement scale is a seven-point Likert Scale, which ranges from 1 to 7 [‘1’ extremely disagree to ‘7’ extremely agree]. This study applies the seven-point Likert Scale to provide wider spread scale values and reduce the possibility of respondent’s bias (Dwivedi, Papazafeiropoulou, Brinkman, & Lal, 2010).

Data Collection Procedure

A total of 850 questionnaires were distributed to the selected schools across the Northern Region of Peninsular Malaysia (Perlis, Kedah, Penang and Perak) using simple random sampling technique. The questionnaire was designed by including the QR code to allow respondents to choose whether to answer on paper or via online. This strategy has successfully speeded up the response time and facilitates the data gathering process. As a result, 719 questionnaires were returned or about 84% response rate, which is considered very high. To elaborate, 441 (61%) respondents answer on paper while 278 (39%) more prefer to answer via online. The data were then gone through the cleaning processes.

Data Cleaning Procedures

There are six main assumptions that are usually considered during the data cleaning, namely missing values, outliers, normality, linearity, homoscedasticity and multicollinearity (Tabachnick & Fidell, 2007). Additionally, this study also examines the Common Method Variance (CMV) as it is occasionally exists in the cross-sectional studies (Juneman, 2013). All of these procedures were done using IBM SPSS Statistics (SPSS) Version 21.

Missing Values

Untreated missing values could cause several problems to data analysis procedures, for examples, data inadequacy and inaccurate result (Gaskin, 2017). Before the analysis of missing values can be done, the accuracy of data entry need to examined, which was done by checking the unexpected values within each variables. As the dataset is confirmed to be free from data entry error, the missing values was treated by using case deletion method (Tabachnick & Fidell, 2007). The case deletion was chosen after considering two justifications; (i) The study achieved a high level of response rate and only a few data points are missing (1.39%), (ii) The cases with the missing values are random subsamples of the whole dataset. During the procedure, 41 cases have been identified as non-users of VLE, therefore have been eliminated from the dataset. Additionally, another ten cases have been diagnosed with the incomplete responses, mainly for the respondents who choose to answer on paper. For those who answer the online questionnaire, each of the items has been set as 'compulsory field' that prevent the respondents from skipping any questions. As the study achieved a high level of response rate, all of the cases with incomplete responses were deleted. In total, 51 case deletions were made at this stage.

Outliers

The outliers among continuous variables usually exist in two forms, univariate and multivariate. Univariate outlier is a case with an extreme value on a single variable, while multivariate outlier is known as an unusual combination of scores in multiple variables that creates statistical anomaly. This study identifies univariate outliers based on unengaged responses (Gaskin, 2017) and standardized scores (z-scores) with the threshold value of ±3.29 (Tabachnick & Fidell, 2007). Unengaged responses occur when the respondent provides same answer for the whole questionnaire, which produced the standard deviation (SD) value of 0.00 for that particular case. In addition, the cases with very large z-scores that are disconnected from other z-scores are also considered as potential outliers. During the analysis, two unengaged responses and one case that has z-score of WL exceeding the threshold value (-3.78) were eliminated from the dataset.

On the other hand, the analysis of multivariate outliers was done using Mahalanobis Distance method. It is one of the measurements for multivariate distance and can be evaluated for each case using the chi-square (X^2) distribution. The most common probability estimate for a case being an outlier is $p < .001$ for the X^2 value, is appropriate with Mahalanobis Distance (Tabachnick & Fidell, 2007). Hence, 22 deletions were made during the procedure, as shown in Table 1.

Table 1: Analysis of Multivariate Outliers

No.	IVs	DV	Deletion	Cases
1.	IQ, SyQ, SeQ, US, NB, U	ITU	15	89, 177, 28, 327, 451, 198, 51, 27, 634, 199, 340, 244, 61, 382, 424
2.	ITU	U	-	-
3	IQ, SyQ, SeQ, NB,U	US	6	396, 93, 560, 408, 559, 586
4.	U,US	NB	1	675

Linearity

Linearity refers to the consistent slope of change between the exogenous and endogenous variables. In this study, the linearity test was done using two methods namely ANOVA test and Ordinary Least Square (OLS) Linear Regression. The relationship between exogenous and endogenous variable is considered as linear when the significance value of ANOVA test is greater than 0.05. However, in occurrence of the opposite result, the OLS Linear Regression should be conducted for crosschecking. The significant value that is lesser than 0.05 is accepted as sufficiently linear (Gaskin, 2017). As shown in Table 2, all the relationships between exogenous and endogenous variables in this study meet the assumption of linearity.

Table 2: Linearity Analysis

Relationship		ANOVA Sig. (> 0.05)	OLS Linear Regression Sig. (<0.05)	Linear Relationship
IQ	ITU	0.250	-	Yes
SyQ	ITU	0.317	-	Yes
SeQ	ITU	0.01	0.00	Yes
US	ITU	0.033	0.00	Yes
NB	ITU	0.017	0.00	Yes
U	ITU	0.037	0.00	Yes
ITU	U	0.653	-	Yes
IQ	US	0.511	-	Yes
SyQ	US	0.518	-	Yes
SeQ	US	0.093	-	Yes
NB	US	0.105	-	Yes
U	US	0.416	-	Yes
U	NB	0.137	-	Yes
US	NB	0.280	-	Yes

Normality

For the distribution normality test, this study applied the analysis of Skewness and Kurtosis, Kolmogorov-Smirnov and Shapiro-Wilk. The initial analysis based on Skewness & Kurtosis indicated that the data are approximately normally distributed or in the range of ±2 (Garson, 2012). However, further analysis of the Kolmogorov-Smirnov and Shapiro-Wilk shown that the significant values are below 0.05, which signify the violation of normality assumption (Table 3). Therefore, this violation calls for the employment of PLS-SEM in the main analysis as it has the ability to handle the standard error of non-normally distributed data (Hair, Ringle, & Sarstedt, 2011).

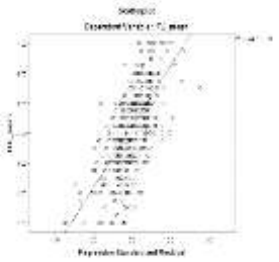
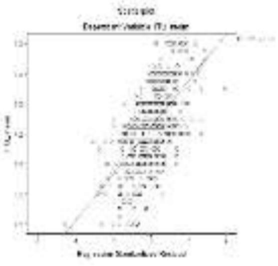
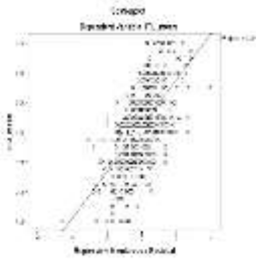
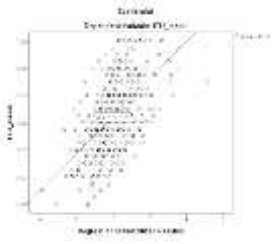
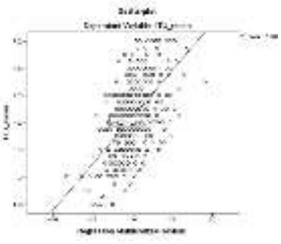
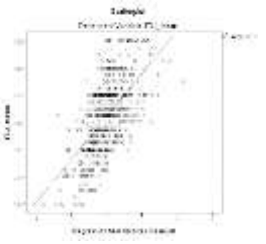
Table 3: Distribution Normality Test

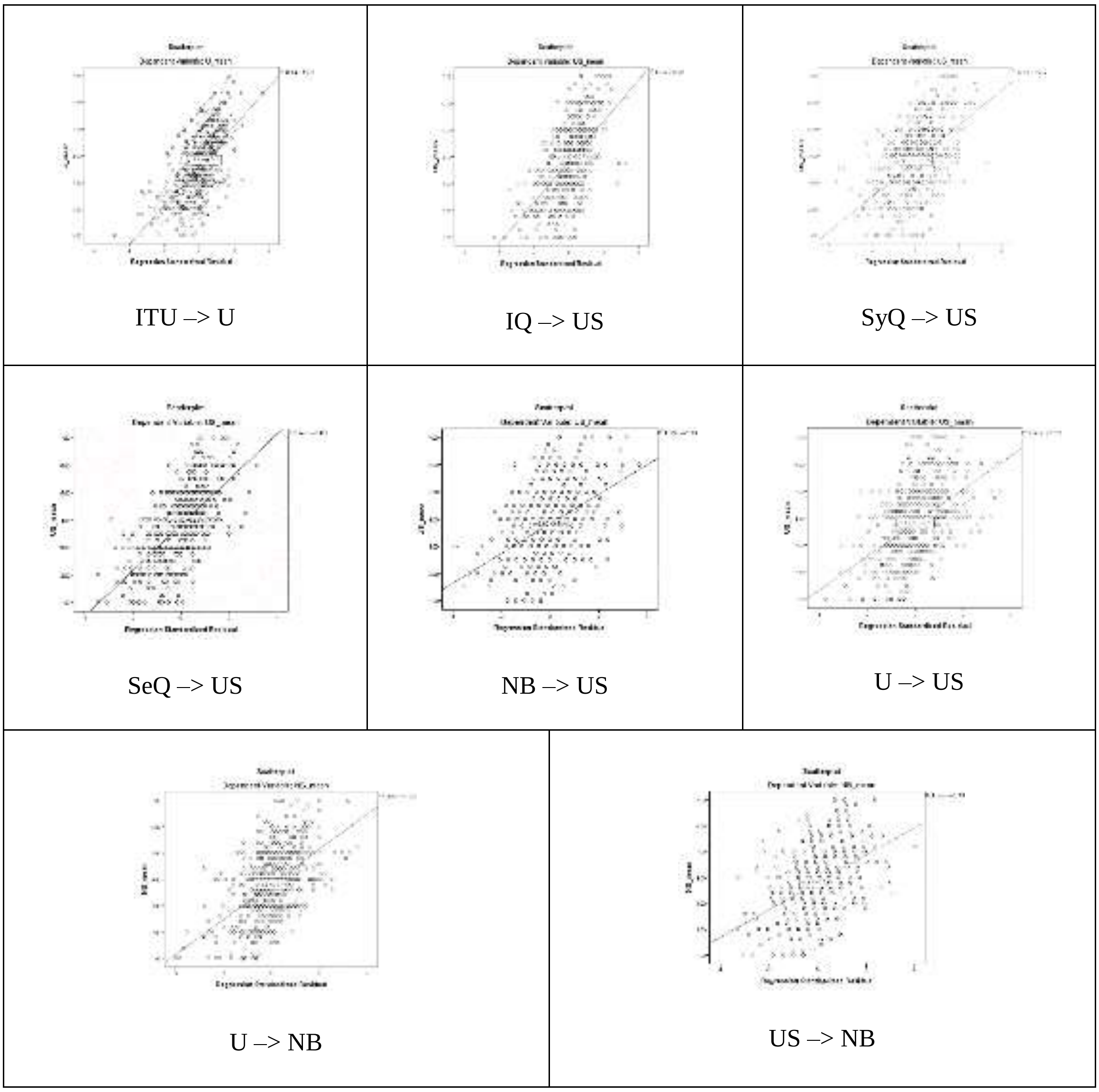
Variable	Skewness	Kurtosis	Kolmogorov-Smirnov (Sig.)	Shapiro-Wilk (Sig.)
IQ	-0.25	0.09	.000	.000
SyQ	-0.18	-0.22	.001	.003
SeQ	-0.05	-0.09	.000	.005
ITU	-0.06	-0.02	.000	.000
U	0.14	-0.32	.001	.007
US	0.02	-0.30	.000	.000
NB	0.001	-0.21	.000	.000
WL	0.11	-0.36	.000	.001

Homoscedasticity

Homoscedasticity exists when the variable's residual exhibits consistent variance across different levels of the variable (Gaskin, 2017). The examination of homoscedasticity can be done using scatterplot analysis, where the equal distance of residuals along the fit line is expected to meet the assumption homoscedasticity. On the contrary, the funnel out shape of scatterplot indicates the existence of heteroscedasticity, which violates the assumption of parametric analysis (Salkind, 2010). As shown in

Figure 2, the scatterplots of all exogenous variables to the endogenous variables are approximately homoscedastic, which signify the good characteristic of data.

 IQ –> ITU	 SyQ –> ITU	 SeQ –> ITU
 US –> ITU	 NB –> ITU	 U –> ITU



ITU –> U

IQ –> US

SyQ –> US

SeQ –> US

NB –> US

U –> US

U –> NB

US –> NB

Figure 2: Homoscedasticity Analysis

Multicollinearity

Multicollinearity appears when the variance of exogenous variables are overlapping with each other and thus not explaining unique variance in the endogenous variable (Gaskin, 2017). For the current study, the Tolerance and Variance Inflation Factor (VIF) values were applied as the indicators in the multicollinearity test. The value of VIF<10 and Tolerance > 0.1 are accepted as the threshold for multicollinearity assumption (Field, 2009). As demonstrated by Table 4, the entire exogenous variables meet the assumption of multicollinearity test, illustrating the inexistence of collinearity issues in the research framework.

Table 4: Multicollinearity Analysis

DV:	ITU		U		US		NB	
IV:	**Tolerance**	**VIF**	**Tolerance**	**VIF**	**Tolerance**	**VIF**	**Tolerance**	**VIF**
IQ	0.24	4.23	-	-	0.24	4.20	-	-
SyQ	0.15	6.55	-	-	0.16	6.22	-	-
SeQ	0.29	3.43	-	-	0.29	3.42	-	-
ITU	-	-	-	-	-	-	-	-
U	0.21	4.77	-	-	0.25	4.05	0.23	4.29
US	0.12	8.57	-	-	-	-	0.23	4.29
NB	0.15	6.55	-	-	0.20	5.01	-	-

Common Method Variance

Common Method Variance (CMV) occurs when the respondents are presumed to have the intention of providing similar answers for different variables, which produced equal variances in both exogenous and endogenous variables. The present evidence shows that CMV can significantly affect items' validity, reliability and covariance between the variables (Podsakoff, MacKenzie, & Podsakoff, 2012). Therefore, it needs to be controlled. One of the methods of detecting CMV is through inter-construct correlation. CMV is expected to exist if the value of inter-construct correlation is above 0.9 (Bagozzi, Yi, & Phillips, 1991). In the current study, only the inter-construct correlation of NB – US is approximately reaching the threshold value (0.91). Nevertheless, this value is considered acceptable as the research framework of the study proposed the recursive relationships between these two variables. Table 5 presents the result obtained from the statistical analysis of CMV in this study.

Table 5: The Analysis of Common Method Variance

	Inter-Construct Correlations							
	IQ	**SyQ**	**SeQ**	**ITU**	**U**	**US**	**NB**	**WL**
IQ	1							
SyQ	0.86	1						
SeQ	0.77	0.78	1					
ITU	0.76	0.79	0.74	1				
U	0.75	0.79	0.73	0.78	1			
US	0.81	0.86	0.77	0.84	0.88	1		
NB	0.79	0.84	0.74	0.84	0.85	0.91	1	
WL	0.26	0.22	0.28	0.22	0.24	0.22	0.18	1

Result and Conclusion

This article discusses the issues that are resolved in the period between data collection and main analysis. This study has successfully acquired 719 raw data during the data collection phase. As can be seen in Table 6, 643 good cases or about 75.6 % of valid response rate were produced after the data cleaning procedures, which meet the requirement of minimum response rate for survey research (Hair, Black, Babin, & Anderson, 2010).Moreover, this amount of cases are valid for further analysis of VLE success, which will be conducted afterwards.

Table 6. Summary of the Data Cleaning Procedures

Cleaning Procedure	Case Deletions / Result
1. Missing Value	51
2. Outliers:	
a. Univariate – 3 deletions	25
b. Multivariate – 22 deletions	
3. Linearity	Linear
4. Normality	Not normal
5. Homoscedasticity	Homoscedastic
6. Multicollinearity	No multicollinearity
7. Common Method Variance	Acceptable
Usable data	**643**

Although these procedures are usually time-consuming and tedious, the careful consideration is necessary to ensure the validity of the main data analysis (Tabachnick & Fidell, 2007). The contamination elements such as incomplete data (missing values) and outliers will negatively affect the accuracy of the finding. In addition, the selection of analysis tool or statistical procedure is heavily relies on the characteristics of data. For example, although Covariance-Based SEM (CB-SEM) is suitable for testing the hypothesized model, it is only practical if the data is normally distributed. Therefore, for non-normal data, the application of PLS-SEM is appropriate. Finally, the assumptions such as linearity, homoscedasticity, multicollinearity and CMV represent the desired characteristics of data. Any violation of these assumptions will requires the treatment of the data. The methods used for this data cleaning and preparation may be applied to another type of studies, especially when dealing with survey data. However, the selection from any of these procedures should also consider the type and purpose of data.

References

Al-Debei, M. M., Jalal, D., & Al-Lozi, E. (2013). Measuring Web Portals Success: A Respecification and Validation of the DeLone and McLean Information Systems Success Model. *International Journal of Business Information Systems*, *14*(1), 96–133. http://doi.org/10.1504/IJBIS.2013.055555

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing Construct Validity in Organizational Research. *Administrative Science Quarterly*, *36*(3), 421–458.

Cheok, M. L., & Wong, S. L. (2016). Frog Virtual Learning Environment for Malaysian Schools: Exploring Teachers' Experience. In J. Zhang et al. (Ed.), *ICT in Education in Global Context* (pp. 201–209). Singapore: Springer Science+Business Media. http://doi.org/10.1007/978-3-662-43927-2

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. *Journal of Management Information Systems*, *19*(4), 9–30. http://doi.org/10.1073/pnas.0914199107

Dwivedi, Y. K., Papazafeiropoulou, A., Brinkman, W. P., & Lal, B. (2010). Examining the Influence of Service Quality and Secondary Influence on the Behavioural Intention to Change Internet Service Provider. *Information Systems Frontiers*, *12*(2), 207–217. http://doi.org/10.1007/s10796-008-9074-7

Field, A. (2009). *Discovering Statistics using SPSS* (3rd ed.). London: SAGE Publications.

Garson, G. D. (2012). *Testing Statistical Assumption*. Asheboro: Statistical Associate Publishing.

Gaskin, J. (2017). Data Screening. Retrieved October 31, 2017, from http://statwiki.kolobkreations.com

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Pearson Prentice Hall.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. http://doi.org/10.2753/MTP1069-6679190202

Johari, H., & Siti Norazlina, K. (2010). Halangan Terhadap Penggunaan Komputer dan ICT di dalam Pengajaran dan Pembelajaran (P&P) di Kalangan Guru di Sekolah Menengah Kebangsaan Luar Bandar di daerah Kulai Jaya, Johor. *Jurnal Pendidik Dan Pendidikan*, *7*(3), 56–67.

Juneman. (2013). Common Method Variance & Bias Dalam Penelitian Psikologis. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *2*(5), 364–381.

Kementerian Kewangan Malaysia. (2014). *Maklum Balas Ke Atas Laporan Ketua Audit Negara 2013 Siri 3*. Putrajaya, Malaysia.

Mohammadi, H. (2015). Factors Affecting the E-Learning Outcomes: An Integration of TAM and IS Success Model. *Telematics and Informatics*, *32*(4), 701–719. http://doi.org/10.1016/j.tele.2015.03.002

Mueller, D., & Strohmeier, S. (2011). Design Characteristics of Virtual Learning Environments: State of Research. *Computers & Education*, *57*(4), 2505–2516. http://doi.org/10.1016/j.compedu.2011.06.017

Nor Azlah, M. J., Fariza, K., Mohd Jaafar, N. A., Khalid, F., Nor Azlah, M. J., Fariza, K., … Khalid, F. (2014). Keberkesanan Kemahiran Komunikasi Di Kalangan Guru Dalam Penggunaan Persekitaran Pembelajaran Maya (Frog VLE). *Pengajaran Sumber Dan Teknologi Maklumat: Impaknya Ke Atas Penyelidikan Dalam Pendidikan 2014*, *11*, 63–69.

Nor Fadzleen, S., Halina, M. D., Haliza, Z., Sa’Don, N. F. B., Dahlan, H. B. M., & Zainal, H. B. (2013). Derivation for Design of Virtual Learning Environment (VLE) framework for Malaysian Schools. In *International Conference on Research and Innovation in Information Systems, ICRIIS* (Vol. 3, pp. 570–575). http://doi.org/10.1109/ICRIIS.2013.6716772

Piccoli, G., Ahmad, R., & Ives, B. (2001). Web-Based Virtual Learning Environments: A Research Framework and a Preliminary Assessment of Effectiveness in Basic IT Skills Training. *MIS Quarterly*, *25*(4), 401–426.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of Method Bias in Social Science Research and Recommendations on How to Control It. *Annual Review of Psychology*, *63*(1), 539–569. http://doi.org/10.1146/annurev-psych-120710-100452

Salkind, N. J. (2010). *Encyclopedia of Research Design*. London: SAGE Publications.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Pearson Education Limited.

Zhu, D. S., Lee, R. Z. C., Kuo, M. J., & Lin, T. S. (2013). A Study on the Continuous Using Intention of Travelling Website. In *2013 IEEE/ACIS 12th International Conference on Computer and Information Science, ICIS 2013 - Proceedings* (pp. 255–261). IEEE Computer Society. http://doi.org/10.1109/ICIS.2013.6607851