Predictive model assessment in PLS-SEM: guidelines for using PLSpredict

Galit Shmueli
National Tsing Hua University, Hsinchu, Taiwan

Marko Sarstedt
Department of Marketing, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany and School of Business and Global Asia in the 21st Century Research Platform, Monash University Malaysia, Jalan Lagoon Selatan, Malaysia

Joseph F. Hair
Department of Marketing, University of South Alabama, Mobile, Alabama, USA

Jun-Hwa Cheah
Department of Management and Marketing, University Putra Malaysia, Serdang, Malaysia

Hiram Ting
Faculty of Hospitality and Tourism Management, University College Sedaya International, Sarawak, Malaysia

Santha Vaithilingam
Monash University Malaysia, Bandar Sunway, Malaysia, and

Christian M. Ringle
Department of Management Sciences and Technology, Hamburg University of Technology (TUHH), Hamburg, Germany and Waikato Management School, University of Waikato, Hamilton, New Zealand

Received 18 February 2019
Revised 18 February 2019
Accepted 26 February 2019

Abstract

Purpose – Partial least squares (PLS) has been introduced as a "causal-predictive" approach to structural equation modeling (SEM), designed to overcome the apparent dichotomy between explanation and prediction. However, while researchers using PLS-SEM routinely stress the predictive nature of their analyses, model evaluation assessment relies exclusively on metrics designed to assess the path model's explanatory power. Recent research has proposed PLSpredict, a holdout sample-based procedure that generates case-level predictions on an item or a construct level. This paper offers guidelines for applying PLSpredict and explains the key choices researchers need to make using the procedure.

Design/methodology/approach – The authors discuss the need for prediction-oriented model evaluations in PLS-SEM and conceptually explain and further advance the PLSpredict method. In addition, they illustrate the PLSpredict procedure's use with a tourism marketing model and provide recommendations on how the results should be interpreted. While the focus of the paper is on the

This research uses of the statistical software SmartPLS (www.smartpls.com). Ringle acknowledges a financial interest in SmartPLS.

0309-0566
DOI 10.1108/EJM-02-2019-0189

PLSpredict procedure, the overarching aim is to encourage the routine prediction-oriented assessment in PLS-SEM analyses.

Findings – The paper advances PLSpredict and offers guidance on how to use this prediction-oriented model evaluation approach. Researchers should routinely consider the assessment of the predictive power of their PLS path models. PLSpredict is a useful and straightforward approach to evaluate the out-of-sample predictive capabilities of PLS path models that researchers can apply in their studies.

Research limitations/implications – Future research should seek to extend PLSpredict's capabilities, for example, by developing more benchmarks for comparing PLS-SEM results and empirically contrasting the earliest antecedent and the direct antecedent approaches to predictive power assessment.

Practical implications – This paper offers clear guidelines for using PLSpredict, which researchers and practitioners should routinely apply as part of their PLS-SEM analyses.

Originality/value – This research substantiates the use of PLSpredict. It provides marketing researchers and practitioners with the knowledge they need to properly assess, report and interpret PLS-SEM results. Thereby, this research contributes to safeguarding the rigor of marketing studies using PLS-SEM.

Keywords PLS-SEM, Partial least squares, Structural equation modeling, PLSpredict, Out-of-sample prediction, Predictive power

Paper type Research paper

1. Introduction

Assessing a statistical model's predictive power is a crucial element of any study. Researchers evaluate theories and the practical relevance of their analyses on the basis of their models' ability to make falsifiable predictions about new observations. According to Hofman *et al.* (2017, p. 486), "historically, this process of prediction-driven explanation has proven uncontroversial in the physical sciences, especially in cases where theories make relatively unambiguous predictions and data are plentiful." In contrast, marketing researchers have generally emphasized prediction less than explanatory modeling, which aims to "test or quantify the underlying causal relationship between effects that can be generalized from the sample to the population of interest" (Shmueli *et al.*, 2016, p. 4553). In other words, marketing researchers' focus is primarily on assessing whether model coefficients are significant, meaningful and in the hypothesized direction, rather than on testing whether a model can predict new cases.

A similar conclusion can be drawn with respect to applications of partial least squares structural equation modeling (PLS-SEM), a widely used regression-based technique in marketing and other social sciences fields, which estimates relationships in path models with latent and manifest variables (Lohmöller, 1989; Wold, 1985; Hair *et al.*, 2017b). Contrary to covariance-based SEM (Jöreskog, 1978; Rigdon, 1998; Diamantopoulos and Siguaw, 2000), which was only designed for explanatory purposes (Sarstedt *et al.*, 2016b), PLS-SEM is a "causal-predictive" method (Jöreskog and Wold, 1982, p. 270). As such, PLS-SEM overcomes the apparent dichotomy between explanation and prediction. While the method maximizes the amount of explained variance of the endogenous constructs embedded in a path model grounded in well-developed causal explanations (Sarstedt *et al.*, 2017), the PLS-SEM results are well suited to generate out-of-sample predictions. Gregor (2006, p. 626) refers to this interplay as explanation and prediction theory, noting that this approach "implies both [an] understanding of [the] underlying causes and prediction, as well as [a] description of [the] theoretical constructs and the relationships among them".

While researchers using PLS-SEM routinely stress the predictive nature of their analyses – as evidenced in numerous reviews of PLS-SEM use across a variety of fields (Ali *et al.*, 2018; Hair *et al.*, 2012a; Hair *et al.*, 2012b; Ringle *et al.*, 2019) – model evaluation relies almost exclusively on metrics designed to assess the path model's explanatory power. Specifically,

most researchers interpret the coefficient of determination (R^2), which assesses the in-sample model fit of the dependent constructs' composite scores, by using the model estimates to predict the case values of the total sample. The R^2 value, however, only assesses a model's explanatory power, but provides no indication of its out-of-sample predictive power in the sense of an ability to predict the values of new cases not included in the estimation process. Assessing a model's out-of-sample predictive power involves estimating the model on a training (analysis) sample and evaluating its predictive performance on data other than the training sample (Cepeda Carrión *et al.*, 2016; Shmueli *et al.*, 2016). Another frequently used metric to assess the model's predictive quality is the Q^2 value (Geisser, 1974; Stone, 1974), which results from the blindfolding procedure (Chin, 1998). Blindfolding omits single data points, but not an entire case, imputes the omitted data points (e.g. by using mean value replacement), and estimates the PLS path model. Using these estimates as input, the blindfolding procedure predicts the omitted data points. As the Q^2 value does not draw on holdout samples, but on single omitted and imputed data points, this metric is a combination of in-sample and out-of-sample prediction without clearly indicating whether the model has good explanatory fit (in an R^2 value sense) or exhibits predictive power (Sarstedt *et al.*, 2017).

This specific focus on metrics to evaluate a model's explanatory power is problematic because for an optimal predictive model may differ from one obtained in an explanatory modeling context. In other words, a well-fitting model designed in an explanatory context may perform poorly in terms of out-of-sample prediction (Shmueli, 2010), thus limiting its practical usefulness. "The critical test of the marketing relevance of explanations is typically provided by the results of actual decision making. The ability of a model to predict with reasonable accuracy the results of one's marketing actions plays a key role" (Steenkamp and Baumgartner, 2000, p. 197). Given the growing concerns about marketing research's practical relevance (Homburg *et al.*, 2015; Reibstein *et al.*, 2009; Lehmann *et al.*, 2011), researchers should include out-of-sample prediction as an integral element of model assessment in PLS-SEM and as a way to assess their model's practical relevance.

Shmueli *et al.* (2016) developed PLSpredict, a holdout-sample-based procedure that generates case-level predictions on an item or a construct level to reap the benefits of predictive model assessment in PLS-SEM. Contrary to standard structural model evaluation metrics such as the R^2 and Q^2, PLSpredict offers a means to assess a model's out-of-sample predictive power (i.e. a model's accuracy when predicting the outcome value of new cases). As PLSpredict has been implemented in standard PLS-SEM software, such as SmartPLS (Ringle *et al.*, 2015) and in R (https://github.com/ISS-Analytics/pls-predict), researchers can use the procedure readily. However, because it has only been developed recently, there are currently very few PLSpredict applications (Felipe *et al.*, 2017; Svensson *et al.*, 2018). Comments in forums such as http://forum.smartpls.com/ suggest that researchers do not know how to interpret PLSpredict results, and recommendations for doing so are vague. In line with this notion, Rasoolimanesh and Ali (2018, p. 243) recently noted that "to date, research has not yet developed clear guidelines for using PLSpredict, which hinders its application".

In this paper, we address this important issue by offering guidelines for applying PLSpredict and by discussing the key choices researchers need to make when initiating the procedure and interpreting its results. We also illustrate the procedure's use with a tourism marketing model and provide recommendations on how the results should be interpreted. While our focus is on the PLSpredict procedure, our overarching aim is to encourage routine prediction-oriented assessment in PLS-SEM analyses.

2. PLSpredict

2.1 Foundations

PLSpredict is based on the concepts of separate training and holdout samples for estimating model parameters, and evaluating a model's predictive power. A training sample is a portion of the overall dataset used to estimate the model parameters (e.g. the path coefficients, indicator weights, and loadings). The remaining part of the dataset not used for model estimation is referred to as the holdout sample (Danks *et al.*, 2017; Hair *et al.*, 2018). To predict the value of a selected dependent construct's indicators, PLSpredict uses the values for the independent constructs' indicators of cases in the holdout sample and applies the model estimates from the training sample to generate prediction of the dependent constructs' indicators (Shmueli *et al.*, 2016)[1]. When computing such predicted values for training sample cases, these are in-sample predictions. In contrast, when computing the predicted values for cases in the holdout sample, these are out-of-sample predictions. A small divergence between the actual and predicted out-of-sample case values suggests that the model has a high predictive power. Conversely, a pronounced divergence between the actual and predicted out-of-sample case values indicates a low predictive power. In addition, researchers can expect the in-sample predictions to be closer to the actual values than the out-of-sample predictions, because the model was estimated using those in-sample training cases. Very large differences between the magnitudes of the in-sample and the out-of-sample deviations between predicted and actual values indicate that the model over-fits the training sample. Such over-fitting generally leads to a low predictive power.

While researchers generally focus on predictions at the indicator level, PLSpredict can also be used to assess a model's predictive power on the grounds of composite scores (Danks *et al.*, 2017). This option is particularly useful when comparing the predictive power of models with different configurations.

Before initiating the PLSpredict procedure, researchers should ensure that all the constructs' measurement models meet the relevant quality standards. In other words, reflectively specified measurement models must exhibit sufficient levels of reliability, convergent validity, and discriminant validity (Hair *et al.*, 2017b; Henseler *et al.*, 2015; Franke and Sarstedt, 2019). Formatively specified measurement models need to be evaluated in terms of their convergent validity, collinearity, and the significance and relevance of the indicator weights (Hair *et al.*, 2017a; Sarstedt *et al.*, 2017).

When running PLSpredict, researchers need to make a series of choices. Most importantly, they need to select a key target construct in the PLS path model for which they want to assess the model's predictive relevance. This construct usually has a reflectively specified measurement model to support the prediction of its items, even though PLS-SEM technically also allows to assess the prediction of a target construct's formatively specified items. Relevant choices when using PLSpredict include:

- the number of folds;
- the number of repetitions; and
- the selection of an adequate prediction statistic to quantify the degree of prediction error.

In the following, we discuss each of these technical PLSpredict choices in greater detail.

2.2 Choosing the number of folds

Rather than using one subset of data to estimate the model (training sample) and predict a holdout sample's case values, PLSpredict is based on the concept of k-fold cross-validation,

in which the overall dataset is split into k equally sized subsets of data (a fold is a subgroup of the total sample). For example, a 5-fold cross-validation splits the total sample into 5 equally sized subsets (groups) of data. PLSpredict then combines k-1 subsets (i.e. 4) into a single training sample to predict the remaining subset, which represents the holdout sample for the first cross-validation run. The cross-validation process is then repeated k times (the folds), with each of the k subsets used exactly once as the holdout sample. Figure 1 illustrates this concept for $k = 5$. In Fold 1, the first subset (Holdout 1) is excluded from the analysis and the model is estimated using the training data, which is made up by combining Folds 2 through 5 into a single sample. The estimates are then used to predict the holdout sample (Holdout 1 predicted). This process is repeated for Fold 2 through Fold 5, yielding predictions for all five holdout samples. For example, Fold 2 becomes the holdout sample and the training sample consists of Fold 1 and Folds 3, 4 and 5. Consequently, each case in every holdout sample has a predicted value based on a model in which that case was not used to estimate the model parameters. The accuracy of these predictions is then summarized in the prediction statistics.

When choosing a value for k, researchers have to ensure that the training sample in a single fold still meets the model's minimum sample size requirements (Kock and Hadaya, 2018). For example, with 200 observations and $k = 5$, each fold's training sample has 160 observations. Hence, the minimum sample to estimate the underlying model must be 160 (or higher). Predictive studies typically set k to 10 (Shmueli *et al.*, 2016). We recommend following this convention as long as the minimum sample size requirements are met. If a k of 10 does not result in the minimum required training sample size, researchers should choose a higher k value to increase the size of the training sample in each cross-validation run.

2.3 Choosing the number of repetitions

When using PLSpredict, researchers have the option of running the algorithm repeatedly. This approach is beneficial if the aim is to predict a new observation using the average of predictions from *multiple estimated models*. In that case, PLSpredict estimates the model r times, generates predictions for each model, and takes the average of these r predictions to

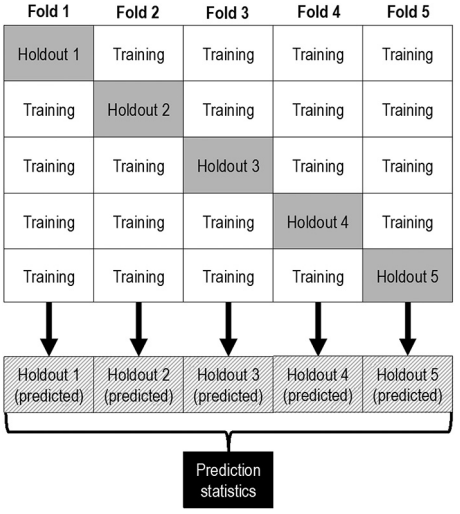


Figure 1. *k*-fold cross-validation with k = 5 folds

predict the value of the new observation. By doing so PLSpredict avoids tapping potentially abnormal solutions that originate from an extreme partitioning of the holdout and the training samples. Although choosing a high value for r increases the estimates' precision, it also increases the algorithm's runtime. Setting r to 10 generally offers a good trade-off between increase in precision and runtime (Witten *et al.*, 2016). Referring to the previous example, the 5-fold cross-validation would be executed 10 times. Repeating the k-fold cross-validation with different random data partitions allows us to compute the prediction errors across all repetitions to ensure a more stable estimate of the predictive performance of the PLS path model.

Alternatively, researchers may set $r = 1$ to mimic how the PLS model will eventually be used to predict a new observation, namely, using a single model (estimated from the entire dataset) for predicting the new observation (Vanwinckelen and Blockeel, 2012). Using a single repetition assures that the results reflect the error variability of predictions generated from that single model. To summarize, setting $r = 10$ is appropriate when the scenario of prediction will be based on an ensemble of r models (i.e. an average prediction from r different estimated models), whereas setting $r = 1$ is adequate when the predictions should be based on a single model.

2.4 Assessing the degree of prediction error

2.4.1 Prediction statistics. Researchers can draw on several prediction statistics to assess their model's predictive power. Popular statistics include the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean squared error (RMSE). The statistics are defined as follows where y_i represents the value of y for observation i ($i = 1,\ldots, n$) and $\hat{y}_i$ is the predicted value for that observation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{3}$$

The MAE measures the average magnitude of the errors in a set of predictions without considering their direction (over or under); it is the average absolute differences between the predictions and the actual observations, with all the individual differences having equal weight. As such, the MAE's values depend on the manifest variables' scaling.

The MAPE offers a more intuitive interpretation by expressing the prediction error in terms of a percentage metric and is therefore independent from the manifest variables' scaling. Although appealing, the MAPE suffers from several limitations, which constrain its practical usefulness. For example, the MAPE is not defined for manifest variables with a zero value, because this would result in a division by zero [equation (2)]. The MAPE is therefore not defined at the composite level, as composite scores are usually close to zero, triggering potential singularity problems. Furthermore, equal errors above the observed

value result in a greater absolute percentage error than those below the observed value (Makridakis, 1993; Tofallis, 2015).

The RMSE is the square root of the average of the squared differences between the predictions and the actual observations. Like the MAE, the RMSE's values depend on the manifest variables' scaling. As the RMSE squares the errors before averaging, the statistic assigns a greater weight to larger errors, which makes it particularly useful when large errors are undesirable. As the latter often applies to predictive modeling, the RMSE rather than other criteria is generally the preferred "default" (Chica and Rand, 2017; Nau, 2016). However, when the distribution of prediction errors is highly non-symmetric, the RMSE may produce an overly pessimistic picture of the model's predictive power. The MAE relaxes this problem, as it weighs all errors equally, thereby being less sensitive to extreme values (Willmott and Matsuura, 2005). This equal weighting also makes the MAE easier to interpret than the RMSE (Pontius *et al.*, 2008).

Sharma *et al.* (2019a) recently evaluated these three statistics' and their related criteria's efficacy in PLS-SEM-based model selection tasks. The results show that the MAE and the RMSE reliably select models that best balance the model fit and the predictive power. Conversely, the MAPE has a pronounced tendency to select incorrectly specified models.

All these measures are single aggregations of a set of errors. Therefore, it is useful to also consider the individual errors by examining their distribution (e.g. using histograms). Such plots convey information about the magnitude and frequency of over- and under-prediction.

2.4.2 Naïve benchmarks. The MAE, MAPE and RMSE are scaled such that smaller values indicate higher predictive power. However, their absolute levels are difficult to interpret in a single model, as their values depend on the scaling of the dependent construct's indicators making any threshold arbitrary. Shmueli *et al.* (2016) suggest using the simple indicator-level average as a naïve benchmark from the training sample. As this procedure is similar to assessing the blindfolding-based Q^2 statistic in PLS-SEM (Stone, 1974; Geisser, 1974), we refer to this naïve benchmark as $Q^2_{predict}$, which uses the mean value of the variables in a training sample as predictions of the variables in the holdout sample. More specifically, $Q^2_{predict}$ equals one minus the quotient of the PLS path model's sum of the squared prediction errors in relation to the mean value's sum of the squared prediction errors. A positive $Q^2_{predict}$ value indicates that the PLS path model's prediction error is smaller than the prediction error given by the (most) naïve benchmark.

While intuitive, the $Q^2_{predict}$ criterion is very simplistic in that it ignores any input information the PLS path model provides. Shmueli *et al.* (2016) therefore propose an alternative benchmark that considers the PLS path model's input layer, while ignoring its concrete structure (e.g. the interrelationships between the constructs). This benchmark uses a linear regression model (LM) to generate predictions for the manifest variables by running a linear regression of each of the dependent construct's indicators on the indicators of the exogenous latent variables in the PLS path model (Evermann and Tate, 2016; McDonald, 1996). This analysis ignores any specified model structure based on measurement and structural theory. Researchers would therefore expect their PLS-SEM-based predictions, which consider the entire model structure (i.e. the measurement and the structural models), to outperform the naïve LM benchmark. That is, the PLS path model's predictive power should be at least equal to that of the LM, with larger improvements demonstrating increasing predictive power (Danks and Ray, 2018).

2.4.3 Guidelines for using PLSpredict. Based on our discussions, Figure 2 shows the process for interpreting PLSpredict results. Recommendations for running PLSpredict (Table I) complement the illustration by listing the rules of thumb for executing and

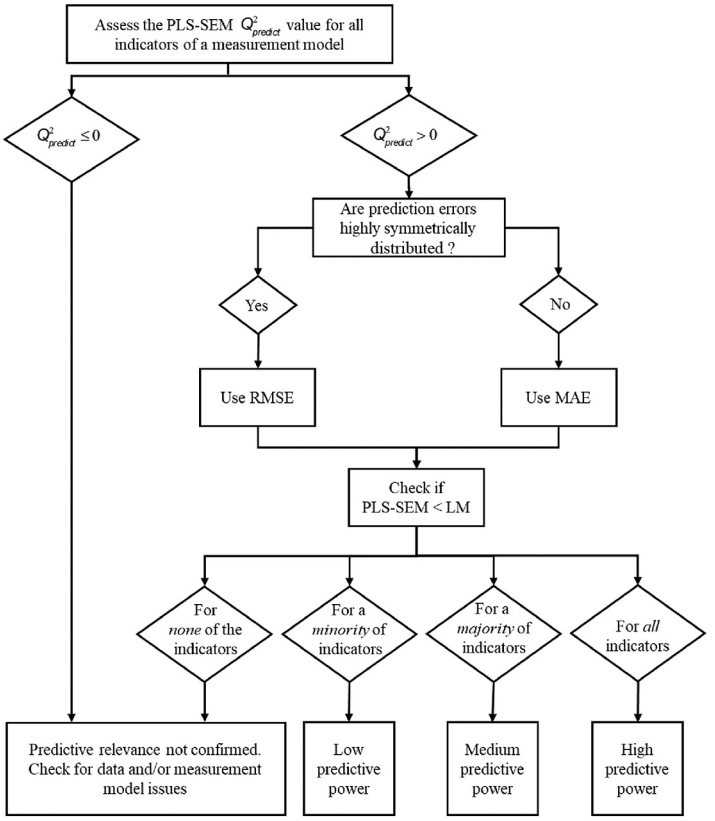


Figure 2.
Guidelines for interpreting PLSpredict results

interpreting PLSpredict. When interpreting PLSpredict results, researchers should generally focus on their model's key endogenous construct, rather than discussing the prediction errors in all of the endogenous constructs' indicators. Once the key target construct has been identified, researchers should first interpret the $Q^2_{predict}$ statistic to ensure that the PLS-SEM-based predictions outperform the most naïve benchmark, followed by the assessment of the prediction statistics (Evermann and Tate, 2016). Because of its conceptual limitations, we do not recommend using the MAPE. Instead, researchers should primarily use the RMSE unless the prediction error distribution is highly non-symmetric. In this case, the MAE is the more appropriate prediction statistic (Table I).

A $Q^2_{predict}$ value of zero or less suggests that the predictive power of the PLS-SEM analysis for that indicator does not even outperform the most naïve benchmark (i.e. the indicator means from the training sample). For those indicators with $Q^2_{predict} > 0$, researchers should next compare the RMSE (or the MAE) values with the naïve LM benchmark. This comparison can have four outcomes:

(1) *PLS-SEM < LM for none of the indicators*: If the PLS-SEM analysis (compared to the LM) yields lower prediction errors in terms of the RMSE (or the MAE) for *none* of the indicators, this indicates that the model lacks predictive power.

Initialization
Use ten folds (i.e., $k = 10$), but ensure that the training sample in a single fold still meets the model's minimum sample size requirements. If not, choose a higher value for k
Use ten repetitions (i.e., $r = 10$) when the aim is to predict a new observation using the average of predictions from multiple estimated models. Alternatively, use one repetition (i.e., $r = 1$) when the predictions should be based on a single model
Interpretation
Assessment of a model's predictive power should primarily rely on one key target construct
To assess the degree of prediction error, use the RMSE unless the prediction error distribution is highly non-symmetric. In this case, the MAE is the more appropriate prediction statistic
Examine each indicator's $Q^2_{predict}$ value from the PLS-SEM analysis. A negative $Q^2_{predict}$ value indicates that the model lacks predictive power
Compare the RMSE (or the MAE) value with the LM value of each indicator. Check if the PLS-SEM analysis (compared to the LM) yields lower prediction errors in terms of RMSE (or MAE) for all (high predictive power), the majority (medium predictive power), the minority (low predictive power), or none of the indicators (lack of predictive power)
Examine the distribution of the prediction errors. PLS-SEM-based residuals should be normally distributed; a left-tailed distribution indicates over-prediction, a right-tailed distribution indicates under-prediction. Also compare the distributions of the prediction errors from PLS-SEM with those from LM. The distributions should correspond closely
Resolving problems in predictive power
Indicators with low predictive power should be analyzed in terms of data issues (e.g. data distribution and outliers) and measurement model issues (e.g. loadings). Consider deleting problematic indicators, but assess the effect of this on the measurement model quality

Table I.
Rules of thumb for running PLSpredict

(2) *PLS-SEM < LM for a minority of the indicators*: If the *minority* of the dependent construct's indicators produces lower PLS-SEM prediction errors compared to the naïve LM benchmark, this indicates that the model has a low predictive power.

(3) *PLS-SEM < LM for a majority of the indicators*: If the *majority* (or the same number) of indicators in the PLS-SEM analysis yields smaller prediction errors compared to the LM, this indicates a medium predictive power.

(4) *PLS-SEM < LM for all indicators*: If *all* indicators in the PLS-SEM analysis have lower RMSE (or MAE) values compared to the naïve LM benchmark, the model has high predictive power.

Examining the prediction error distribution not only offers guidance regarding the choice of the best prediction statistic (i.e. MAE in case of highly non-symmetric errors, RMSE else; Table I), but also provides evidence of systematic biases. Generally, the errors should follow a normal distribution – substantial deviations from normality suggest that the PLS path model does not have sufficient predictive power (Danks *et al.*, 2017). Further checking of the distribution's shape allows for substantiating whether the model systematically over- or under-predicts the case values. Negative errors, indicated by a long left tail in the residual distribution, suggest an over-prediction, whereas positive errors, indicated by a long right tail in the residual distribution, indicate an under-prediction (Danks and Ray, 2018). Researchers can also compare the PLS-SEM and the LM prediction errors. The two prediction error distributions should generally correspond closely.

If the PLSpredict identifies one or more indicators with a low predictive power, researchers should carefully explore potential explanations. These include:

- data issues; and
- measurement model issues.

In terms of data issues, predictions can be off because of the prediction error's bias and/or variance. For example, if an indicator has a very large variance (across observations), this means that the prediction for observations far from the mean will suffer. Specifically, using single items to measure abstract concepts has a deteriorating effect on a model's predictive power (Diamantopoulos *et al.*, 2012; Sarstedt *et al.*, 2016a; Salzberger *et al.*, 2016). In the case of single-item measures of abstract concepts, respondents are asked to "automatically consider different aspects of the construct" (Fuchs and Diamantopoulos, 2009, p. 204). But asking respondents to do this is likely to increase the error variance, thus negatively influencing a model's capability to predict observations in a training sample. Similarly, if the data have a U-shaped distribution and the prediction is the mean (i.e. the middle of the U), the predictions for most observations are wrong, although they may, on average, be correct. Reconsidering the outlier treatment (i.e. criteria chosen for removing outliers) or transforming the problematic indicator (Mooi *et al.*, 2018, Chapter 5) may increase its predictive power.

Low predictive power could also be because of measurement model issues. For example, an indicator loading could be low, but still be sufficiently high for the construct to meet common reliability and convergent validity thresholds (see Exhibit 4.4 in Hair *et al.*, 2017b). But if the same indicator lacks predictive power, researchers should carefully consider removing it, even if measurement theory supports its inclusion. This conclusion (indicator removal) holds particularly in situations in which the analysis's primary objective is prediction, or when a negative $Q^2_{predict}$ value points to a very low predictive power. However, researchers should not mechanically remove corresponding indicators, but carefully evaluate indicator deletion's effect on the construct's reliability, content validity, convergent validity, and discriminant validity (Hair *et al.*, 2017b; Sarstedt *et al.*, 2017).

2.4.4 Using PLSpredict for model comparisons. When studying some phenomenon of interest, researchers often face alternative explanations, which give rise to several models with different (or additional) antecedents and/or model relationships, all of which are plausible within the realm of the theoretical framework(s) under consideration. Such alternative competing models may also arise when researchers seek to build conceptual bridges across related streams of inquiry to provide a holistic understanding of the phenomenon (Sharma *et al.*, 2019a).

PLSpredict facilitates the empirical comparison of competing models with the same endogenous dependent variable in terms of their predictive power. PLSpredict produces case-specific predictions on the composite level, which researchers can use as input to compare competing models from an out-of-sample prediction perspective. Specifically, researchers should establish a set of competing models and choose the model that minimizes the relevant out-of-sample error statistic (e.g. MAE or RMSE).

However, if the aim is to balance the prediction and the explanation in model assessment by identifying, from among a set of potentially reasonable models, a model that "provides predictions and has both testable propositions and causal explanations" (Gregor, 2006, p. 620), researchers should compare models by using information-theoretic model selection criteria (McQuarrie and Tsai, 1998). These criteria optimize predictive accuracy by striking a balance between model fit and model complexity to avoid over-fitting. Importantly, contrary to PLSpredict, information-theoretic model selection criteria do *not* require the use of a holdout sample. Instead, they avoid over-fitting the data used for the model estimation by including a

penalty in the model fit. This approach is especially useful when the dataset is very small or the data partitioning is problematic. Research has produced a range of criteria, which differ in their statistical underpinnings and in terms of their efficacy for model comparisons in different contexts (Burnham and Anderson, 2002). Sharma *et al.*'s (2019a, 2019b) simulation studies found that the Bayesian Information Criterion (Schwarz, 1978) and the Geweke–Meese criterion (Geweke and Meese, 1981a) achieved the best balance between theoretical consistency and high predictive power, even in the absence of a holdout sample.

3. Empirical example

3.1 Study design and data

We draw on Amaro and Duarte's (2015) model of consumers' intentions to purchase journeys online to illustrate the use of the PLSpredict procedure with empirical data. Grounded in the theory of reasoned action (Ajzen and Fishbein, 1980), the model's goal is to explain the effects of trust (*TRUST*) and communicability (*COMM*) on attitude (*ATT*), and finally on the intention to purchase a journey online (*IPTO*). The model also includes the following five antecedent constructs of *TRUST* and *COMM*, which express the perceived relative advantages:

(1) convenience (*CONV*);
(2) time saving (*TS*);
(3) financial advantages (*FA*);
(4) the product variety available by purchasing journeys online (*PV*); and
(5) the enjoyment of purchasing journeys online (*ENJOY*).

Figure 3 illustrates the path model under consideration. All the constructs draw on reflective measurement models, except for *ATT*, which is measured with a single item. All the items were measured on a seven-point Likert scale; the Appendix provides an overview of all the constructs and their measurement items.

The survey design followed a sequence of steps, including a pre-test (three postgraduate students and five adult volunteers who regularly purchase journeys online) and a pilot test with 30 respondents to identify problematic items and further improve the survey. The online survey-based data collection resulted in 385 responses of online journey purchasers who regularly visit popular travel websites (e.g. Booking, TripAdvisor, Expedia, Hotels, Travelocity and Airbnb). We had to discard 35 responses because of straight lining (Sarstedt and Mooi, 2019, Chapter 5). The final dataset of 350 observations primarily consists of female Malay journey purchasers between 21 and 30 years and who have a

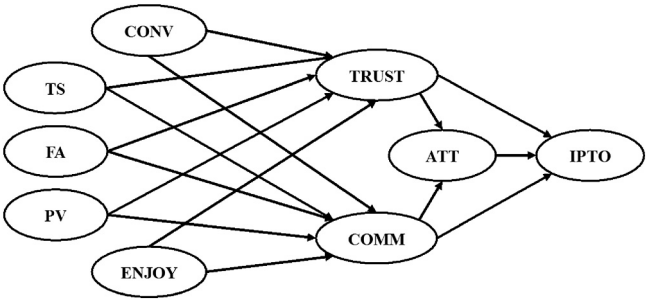


Figure 3.
Path model

monthly income of RM 2,001-4,000 (approx €425-850). Most of the respondents are annual online journey purchasers who book family trips. Table AI shows a detailed breakdown of the sample characteristics.

3.2 Data analysis

3.2.1 Assessment of the measurement models. The analysis begins with an assessment of the measurement models (Hair *et al.*, 2017b, Chapter 4; Chin, 2010). The results show that all of the reflectively measured constructs' measures are reliable and valid (Table II). More specifically, all the loadings exceed the threshold value of 0.708, the average variance extracted (AVE) is higher than the critical value of 0.5, and all the construct reliabilities (i.e. Cronbach's alpha, the coefficients ρ_A, and the composite reliability ρ_C) have values above

Table II. Assessment of convergent validity and internal consistency reliability

Item	Loading	Cronbach's alpha	ρ_A	ρ_C	AVE
Attitude (ATT)					
att	–	–	–	–	–
Communicability (COMM)					
comm1	0.846	0.813	0.815	0.889	0.727
comm2	0.850				
comm3	0.863				
Convenience (CONV)					
conv1	0.877	0.839	0.849	0.903	0.756
conv2	0.882				
conv3	0.848				
Enjoyment (ENJOY)					
enjoy1	0.944	0.884	0.885	0.945	0.896
enjoy2	0.949				
Financial advantages (FA)					
fa1	0.870	0.828	0.83	0.897	0.744
fa2	0.876				
fa3	0.840				
Intention to purchase travel online (IPTO)					
ipto1	0.920	0.760	0.785	0.892	0.805
Ipto2	0.873				
Product variety (PV)					
pv1	0.875	0.770	0.803	0.865	0.682
pv2	0.788				
pv3	0.813				
Time saving (TS)					
ts1	0.881	0.848	0.849	0.908	0.767
ts2	0.876				
ts3	0.869				
Trust (TRUST)					
trust1	0.737	0.855	0.868	0.896	0.632
trust2	0.832				
trust3	0.838				
trust4	0.797				
trust5	0.767				

0.7 (Sarstedt *et al.*, 2017). Finally, the discriminant validity assessment, based on Henseler *et al.*'s (2015) heterotrait-monotrait ratio of correlations (HTMT) measure, shows that all the HTMT values are significantly lower than 0.85 (Franke and Sarstedt, 2019), thus supporting the measures' discriminant validity (Table III).

3.2.2 Assessment of the structural model

3.2.2.1 Collinearity. In line with the structural model assessment procedure outlined in Hair *et al.* (2017a, Chapter 6), we first assess the structural model for collinearity issues by examining the variance inflation factor (VIF) values of all the predictor constructs in the model. As all the VIF values are below the more conservative threshold of 3.3 (Diamantopoulos and Siguaw, 2006), we conclude that collinearity is not at critical levels (Table IV; Model 1).

3.2.2.2 Significance and relevance of the path coefficients. The results of the bootstrapping procedure with 10,000 samples and using the no sign changes option (Streukens and Leroi-Werelds, 2016) reveal that most of the structural model relationships are significant (Table IV; Model 1). Specifically, we find that *COMM* has a significant and meaningful effect on *ATT* (0.457, $p < 0.01$) and *IPTO* (0.362, $p < 0.01$), whereas *TRUST*'s impact on these two constructs is much less pronounced (*ATT*: 0.221, $p < 0.01$; *IPTO*: 0.051, $p > 0.10$). Analyzing the impact of *TRUST* and *COMM's* antecedent constructs, we find that *CONV* has a significant effect on both *TRUST* (0.187, $p < 0.01$) and *COMM* (0.253, $p < 0.01$). The effects of *FA* and *ENJOY* are, however, only significant with regard to *TRUST* (*FA*: 0.239, $p < 0.01$; *ENJOY*: 0.239, $p < 0.01$).

3.2.2.3 In-sample model fit. To assess the model's in-sample fit, we first consider the R^2 (Table IV; Model 1). We find that all the endogenous constructs have R^2 values of around 0.3. This level is clearly lower than in Amaro and Duarte's (2015) study, which considered a more complex model with additional antecedent constructs. While the model's in-sample model fit is rather small according to absolute standards (Hair *et al.*, 2017b, Chapter 6), we consider it acceptable for this study in light of the model's complexity.

3.2.2.4 Out-of-sample predictive power. We use PLSpredict with 10 folds and one repetition to mimic how the PLS model will eventually be used to predict a new observation, rather than using the average of across multiple models[2]. To illustrate the interpretation, we focus our analysis on the model's key target construct *IPTO* (Table I), but also report the prediction statistics of all the other endogenous constructs' indicators.

In a first step, we find that all the endogenous constructs' indicators outperform the most naïve benchmark (i.e. the training sample's indicator means), as all the indicators yield $Q^2_{predict}$ values above 0 (Table V). Next, we analyze the prediction errors in greater detail to identify the relevant prediction statistic. The plots in Figures 4 and 5 (left panels) suggest that the PLS-SEM errors are not normally distributed, which the results of the Shapiro–Wilk test (*ipto1*: $z = 4.720$, $p < 0.01$; *ipto2*: $z = 5.208$, $p < 0.01$) also support. Nevertheless, the visual inspection of the prediction errors suggests that the distribution is not highly non-symmetric. Hence, we base our predictive power assessment on the RMSE; note, however, that the MAE analysis does not lead to substantially different findings in our example.

In a first step, we can evaluate the indicator-level RMSE in the scale of the indicator. By design, the two indicators of *IPTO* are measured on a seven-point scale and have an average RMSE of 0.979. We can thus say that on average 68 per cent of prediction errors will fall within approximately two points of the seven-point scale. For example, if the true value is four, 68 per cent of the predictions will fall between 3.021 (4-1·RMSE) and 4.979 (4 + 1·RMSE) and 95 per cent of all predictions will fall between 2.042 (4-2·RMSE) and 5.958 (4 + 2·RMSE). This range represents nearly the full range of the indicators' measurement scale and one needs to consider if this is acceptable given the context (Danks and Ray, 2018).

Table III.
Assessment of discriminant validity using the heterotrait-monotrait ratio of correlations criterion

Construct	1	2	3	4	5	6	7	8	9
1. ATT									
2. COMM	0.579 [0.478; 0.666]								
3. CONV	0.533 [0.453; 0.610]	0.591 [0.490; 0.685]							
4. ENJOY	0.364 [0.265; 0.453]	0.400 [0.273; 0.511]	0.500 [0.362; 0.622]						
5. FA	0.439 [0.344; 0.523]	0.452 [0.325; 0.563]	0.677 [0.571; 0.770]	0.650 [0.541; 0.745]					
6. IPTO	0.570 [0.452; 0.672]	0.665 [0.553; 0.767]	0.593 [0.498; 0.683]	0.436 [0.309; 0.555]	0.467 [0.338; 0.587]				
7. PV	0.348 [0.233; 0.448]	0.452 [0.323; 0.566]	0.570 [0.423; 0.689]	0.568 [0.430; 0.680]	0.631 [0.521; 0.727]	0.406 [0.276; 0.530]			
8. TS	0.499 [0.410; 0.581]	0.612 [0.502; 0.710]	0.755 [0.664; 0.837]	0.615 [0.485; 0.725]	0.667 [0.574; 0.749]	0.632 [0.531; 0.733]	0.649 [0.528; 0.751]		
9. TRUST	0.377 [0.281; 0.470]	0.333 [0.226; 0.443]	0.492 [0.386; 0.584]	0.517 [0.412; 0.620]	0.557 [0.449; 0.656]	0.305 [0.191; 0.417]	0.340 [0.217; 0.456]	0.451 [0.346; 0.544]	

Note: Numbers in brackets represent the 95% bias-corrected and accelerated confidence intervals derived from bootstrapping with 10,000 samples (no sign change option)

Model	Relationship	Std beta	p-value	95% BCa confidence interval LB	UB	VIF	R^2	Q^2
1	CONV → TRUST	0.187	0.001	0.064	0.311	1.923	0.305	0.179
	TS → TRUST	0.043	0.263	−0.091	0.178	2.130		
	FA → TRUST	0.239	0.000	0.113	0.360	1.918		
	PV → TRUST	−0.049	0.211	−0.171	0.070	1.592		
	ENJOY → TRUST	0.239	0.000	0.108	0.370	1.662		
	CONV → COMM	0.253	0.000	0.126	0.375	1.923	0.312	0.210
	TS → COMM	0.280	0.000	0.144	0.411	2.130		
	FA → COMM	0.013	0.416	−0.109	0.146	1.918		
	PV → COMM	0.073	0.094	−0.033	0.177	1.592		
	ENJOY → COMM	0.040	0.253	−0.078	0.160	1.662		
	TRUST → ATT	0.221	0.000	0.135	0.306	1.096	0.317	0.308
	COMM → ATT	0.457	0.000	0.358	0.544	1.096		
	TRUST → IPTO	0.051	0.139	−0.041	0.140	1.168	0.352	0.266
	ATT → IPTO	0.293	0.000	0.155	0.416	1.464		
	COMM → IPTO	0.362	0.000	0.251	0.475	1.401		
2	CONV → TRUST	0.187	0.001	0.061	0.301	1.923	0.304	0.179
	TS → TRUST	0.041	0.273	−0.092	0.174	2.130		
	FA → TRUST	0.239	0.000	0.114	0.366	1.918		
	PV → TRUST	−0.050	0.204	−0.178	0.063	1.593		
	ENJOY → TRUST	0.240	0.000	0.108	0.369	1.662		
	CONV → COMM	0.253	0.000	0.125	0.382	1.923	0.312	0.209
	TS → COMM	0.280	0.000	0.132	0.404	2.130		
	FA → COMM	0.013	0.419	−0.114	0.144	1.918		
	PV → COMM	0.073	0.095	−0.040	0.175	1.593		
	ENJOY → COMM	0.040	0.253	−0.084	0.158	1.662		
	TRUST → ATT	0.221	0.000	0.132	0.307	1.094	0.317	0.308
	COMM → ATT	0.457	0.000	0.353	0.545	1.094		
	ATT → IPTO	0.499	0.000	0.395	0.589	1.000	0.249	0.191
3	CONV → TRUST	0.185	0.001	0.065	0.305	1.921	0.305	0.179
	TS → TRUST	0.043	0.264	−0.088	0.178	2.132		
	FA → TRUST	0.241	0.000	0.113	0.360	1.916		
	PV → TRUST	−0.049	0.212	−0.173	0.066	1.596		
	ENJOY → TRUST	0.238	0.000	0.098	0.364	1.663		
	CONV → COMM	0.253	0.000	0.122	0.370	1.921	0.312	0.210
	TS → COMM	0.280	0.000	0.140	0.406	2.132		
	FA → COMM	0.012	0.425	−0.111	0.144	1.916		
	PV → COMM	0.073	0.090	−0.039	0.178	1.596		
	ENJOY → COMM	0.040	0.254	−0.081	0.157	1.663		
	CONV → ATT	0.183	0.001	0.066	0.306	2.058	0.370	0.345
	TS → ATT	0.093	0.066	−0.029	0.220	2.248		
	FA → ATT	0.062	0.164	−0.068	0.193	1.999		
	PV → ATT	−0.011	0.418	−0.119	0.098	1.608		
	ENJOY → ATT	0.026	0.330	−0.094	0.139	1.746		
	TRUST → ATT	0.106	0.020	0.005	0.206	1.441		
	COMM → ATT	0.325	0.000	0.211	0.428	1.456		
	TRUST → IPTO	0.051	0.136	−0.038	0.143	1.168	0.352	0.266
	ATT → IPTO	0.293	0.000	0.171	0.405	1.464		
	COMM → IPTO	0.362	0.000	0.248	0.477	1.401		

Table IV. Assessment of the structural model

Comparing the RMSE values from the PLS-SEM analysis with the naïve LM benchmark (Table V), we find that the PLS-SEM analysis produces lower prediction errors for all the indicators. For example, when using PLS-SEM to estimate the model, indicators *ipto1*, and *ipto2* have RMSE values of 0.983, and 0.974, whereas the LM produces RMSE values of 0.986 and 0.975 for these indicators. The differences are more pronounced for indicators *comm1*, *comm2*, and *comm3*, which have PLS-SEM-based RMSE values of 0.896, 0.973 and 0.885, compared to 0.928, 1.013 and 0.919 in the LM.

The longer left tails of the prediction error distributions (Figures 4 and 5) suggest that the PLS path model tends to slightly over-predict the case values (Danks and Ray, 2018). Nevertheless, because the indicators' mean bias is very small (*ipto1* = 0.0007; *ipto2* = 0.0007), we conclude that the model predicts *ipto1* and *ipto2* sufficiently well.

Figures 4 and 5 also contrast the distribution of the *ipto1* and *ipto2* prediction errors from PLS-SEM (left panels) with those from the LM (right panels). The kernel density functions of

Table V.
PLSpredict assessment of manifest variables (original model)

Item	PLS-SEM RMSE	PLS-SEM $Q^2_{predict}$	LM RMSE	PLS-SEM - LM RMSE
att	0.923	0.250	0.935	−0.012
comm1	0.896	0.186	0.928	−0.032
comm2	0.973	0.203	1.013	−0.040
comm3	0.885	0.239	0.919	−0.034
ipto1	0.983	0.220	0.986	−0.003
ipto2	0.974	0.179	0.975	−0.001
trust1	1.187	0.122	1.204	−0.017
trust2	0.959	0.217	0.976	−0.017
trust3	0.996	0.201	1.018	−0.022
trust4	0.978	0.215	0.984	−0.006
trust5	1.118	0.116	1.134	−0.016

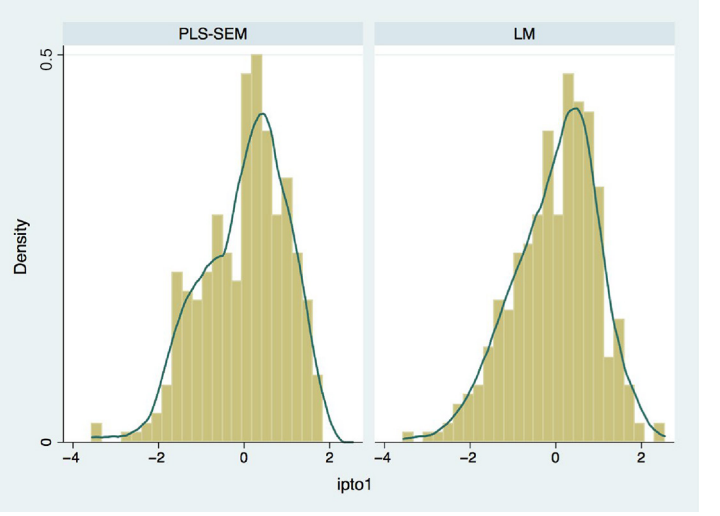


Figure 4.
Residual plot of *ipto1*

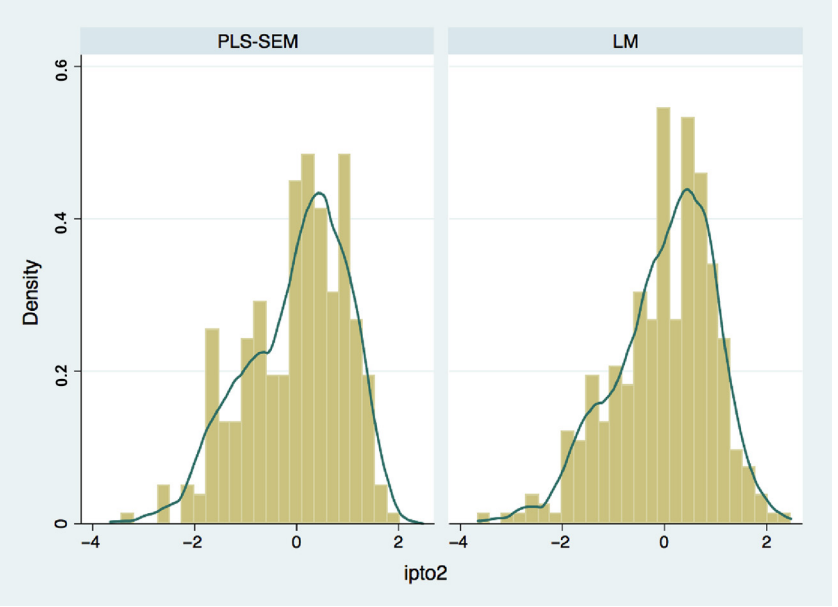
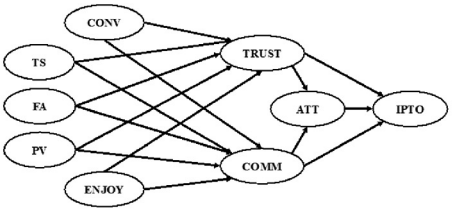


Figure 5.
Residual plot of *ipto2*

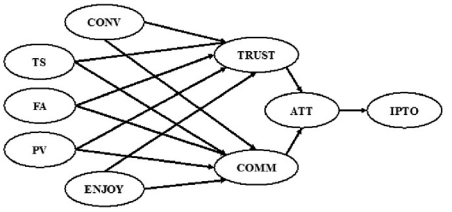
the two approaches generally correspond closely, suggesting that PLS-SEM and the LM produce similar prediction errors. Divergences occur in negative errors between −0.5 and −1.5, where PLS-SEM produces prediction errors that tend to follow a multimodal distribution, while the LM does not. This result is expected, as the PLS-SEM analysis considers the mediating role of the *TRUST*, *COMM*, and *ATT* constructs, whereas this is not the case for LM. Researchers could identify those observations whose mediating constructs entail a greater prediction bias as evidenced by the significant deviations in the two curves. Identifying characteristics that these observations share allows for advancing the existing theory and modifying the path model accordingly.

3.2.2.5 Model comparison. In a final step, we use the PLSpredict results on the construct level to compare the original Model 1 with two theoretically justified alternative models for explaining consumers' intention to purchase journeys online (Figure 6)[3]. The aim of this analysis is to assess the original model's robustness (Sarstedt *et al.*, 2019). Specifically, Model 2 is a simplified version of Model 1 in that only *ATT* influences *IPTO* directly. In contrast, Model 3 is a more complex variant of the original model in which the antecedent constructs also influence *ATT* directly.

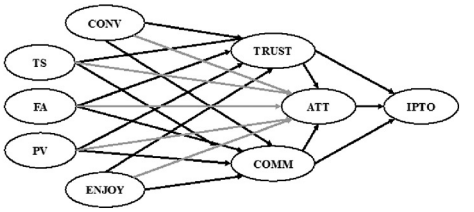
In an initial step, we analyze the prediction errors of the latent variable scores and find that the prediction error distribution of the *IPTO* construct's scores is almost symmetric. Hence, we base our model selection decision on the RMSE statistic. Comparing the three model configurations in terms of RMSE values reveals that the original Model 1 and the parsimonious Model 2 variant clearly outperform the more complex Model 3 in terms of latent variable score-based prediction errors (Table VI)[4]. This result is consistent with Myung's (2000) notion that a more complex model, such as Model 3, often generalizes poorly to new datasets, because it could identify spurious patterns in a sample and hence overfits the original data by absorbing random error. In contrast, models with fewer parameters, such as Model 2, have a better chance of being scientifically replicable, explainable, and



Model 1 (original model)



Model 2



Model 3

Figure 6.
Alternative path models

Model	Latent variable	RMSE	MAE
1	ATT	*0.356*	*0.278*
	COMM	0.543	0.419
	ITPO	0.321	0.249
	TRUST	*0.541*	*0.416*
2	ATT	0.357	0.279
	COMM	*0.538*	*0.414*
	ITPO	*0.177*	*0.138*
	TRUST	0.542	0.417
3	ATT	0.532	0.416
	COMM	0.540	0.416
	ITPO	0.367	0.283
	TRUST	0.543	0.418

Note: Minimum values per construct printed in italic

Table VI.
Model comparison using PLSpredict

exhibiting higher predictive abilities (Shmueli and Koppius, 2011; Bentler and Mooijaart, 1989). Model 2 stands out particularly because of its low prediction error in the *ITPO* construct. As *IPTO* is the final target construct and because Model 2 is more parsimonious, we would prefer this model over the original version as proposed by Amaro and Duarte (2015).

4. Discussion

PLS-SEM's mainstay is creating explanatory models with acceptable predictive power. By maximizing the amount of explained variance of the endogenous constructs embedded in a hypothesized path model, PLS-SEM discloses the mechanisms by which the prediction is being produced. This feature differentiates PLS-SEM from the various machine learning tools designed for prediction purposes, which, in contrast to PLS-SEM, usually do not build on a model of theoretically established causal relationships between the variables of interest.

Instead, PLS-SEM "aims to maintain interpretability while engaging in predictive modeling" (Shmueli *et al.*, 2016, p. 4552).

To reap the benefits of PLS-SEM's predictive capabilities, however, researchers must have access to tools that enable them to generate in-sample and out-of-sample predictions from their model, which involves estimating the model on a training sample that excludes the holdout sample cases to be predicted. Shmueli *et al.*'s (2016) PLSpredict procedure meets this need by translating *k*-fold cross validation – a procedure well-known from predictive analytics – into a PLS-SEM framework.

Despite prediction and predictive assessment's obvious relevance for the field, PLSpredict applications are still scarce, probably because of its recency and the need for straightforward guidance and illustration of its usage. This study offers guidelines for using PLSpredict to increase researchers' awareness of the procedure and to offer clear guidance on how to run and interpret its results. Specifically:

- we describe how to initiate a PLSpredict analysis and interpret its results;
- introduce benchmarks, which allow for assessing a model's out-of-sample predictive power; and
- discuss ways to handle indicators with a low predictive power.

Our illustration in the context of a model of consumers' intentions to purchase journeys online illustrates our guidelines.

PLSpredict enables researchers to address long-standing calls for a stronger focus on predictive model assessment, most notably a model's out-of-sample predictive power (Rigdon, 2012; Sarstedt *et al.*, 2014). However, the procedure also offers value in other contexts, such as scale development and index construction studies. While assessing the predictive power of a newly developed scale or index is an important component of its validation process (Sethi and King, 1991; Diamantopoulos and Siguaw, 2006), PLS-SEM-based studies generally lack such assessments (Musa *et al.*, 2018; Oliveira *et al.*, 2018; Saayman *et al.*, 2018).

The results of our model comparison example illustrate the interplay between model complexity and predictive power. Increased model complexity could give an impression of improved explanatory power in terms of R^2, but it could come at the cost of its generalizability (Myung, 2000). Simultaneously, an overly simplistic model could produce PLS-SEM-based prediction errors that are nearly identical to the naïve LM benchmark. PLSpredict's separation of training and holdout samples allows researchers to strike a balance between these two extremes. In addition, comparing models in terms of composite score-based prediction errors (e.g. using the RMSE or the MAE statistic) enables researchers to identify a parsimonious model that is more likely to exhibit higher predictive power and generalize to other samples. Furthermore, a path model with a weak explanatory power, but a strong predictive performance, might point to the need to extend the existing theory or develop a new one. That is, a researcher could investigate *why* predictions are so accurate and whether their underlying correlations could be linked to a causal theory, thus heralding a theory building process (Gregor, 2006). Further assessment by means of information-theoretic model evaluation criteria could guide researchers' effort to balance prediction and explanation. Sharma *et al.*'s (2019a, 2019b) recent simulation studies identified the Bayesian Information Criterion (Schwarz, 1978) and the Geweke–Meese criterion (Geweke and Meese, 1981b) as particularly effective in this regard and for selecting a model from a set of potentially viable models.

While our study sets the stage for the routine use of PLSpredict, future research should seek to extend its capabilities. For example, researchers should develop more benchmarks

for comparing PLS-SEM results. Such benchmarks could consider different model layers that are known (e.g. the relationship between *TRUST* and *COMM* and their five antecedent constructs), or assume other functional forms in the structural model relationships. Danks and Ray (2018) differentiate between the earliest antecedent (EA) and the direct antecedents (DA) approaches to predictive power assessment. The EA approach treats any mediators as intervening variables, and predicts the composite score from only earliest exogenous antecedents, On the contrary, the DA approach treats mediators as purely exogenous constructs, thus, dropping earlier antecedent constructs[5]. As Danks and Ray (2018, p. 40) note, "both techniques have shortcomings in that they necessarily ignore some part of the measurement model – the indicators and weights of the mediator in the case of the EA approach, or the indicators and weights of earlier antecedents in the case of the DA approach." Future research should empirically contrast both approaches to predictive power assessment and derive guidelines for their use.

Notes

1. Note that this approach is different from jackknifing, which uses the predictions of the holdout sample to estimate the bias and variance of a statistic of interest (Cameron and Trivedi, 2005).
2. Results from using PLSpredict with ten repetitions parallel those presented here.
3. Note that model comparisons usually come *before* measurement model evaluation and predictive power assessment (Sharma *et al.*, 2019b). Alternatively, model comparisons are sometimes used as robustness checks to ascertain the original model's stability. Our illustration follows the latter approach.
4. Table IV shows each model's structural model estimates.
5. SmartPLS 3 version 3.2.8 (Ringle *et al.*, 2015) has implemented the PLSpredict procedure based on the EA approach. Future releases will also include, as an alternative, the DA approach option.

References

Ajzen, I. and Fishbein, M. (1980), *Understanding Attitudes and Predicting Social Behavior*, Prentice Hall, Englewood Cliffs, NJ.

Ali, F., Rasoolimanesh, S.M., Sarstedt, M., Ringle, C.M. and Ryu, K. (2018), "An assessment of the use of partial least squares structural equation modeling (PLS-SEM) in hospitality research", *International Journal of Contemporary Hospitality Management*, Vol. 30 No. 1, pp. 514-538.

Amaro, S. and Duarte, P. (2015), "An integrative model of consumers' intentions to purchase travel online", *Tourism Management*, Vol. 46, pp. 64-79.

Bentler, P.M. and Mooijaart, A. (1989), "Choice of structural model via parsimony: a rationale based on precision", *Psychological Bulletin*, Vol. 106 No. 2, pp. 315-317.

Bigné, E., Sanz, S., Ruiz, C. and Aldás, J. (2010), "Why some internet users don't buy air tickets online", in Gretzel, U., Law, R. and Fuchs, M. (Eds), *Information and Communication Technologies in Tourism*, Springer, Vienna, pp. 209-221.

Burnham, K.P. and Anderson, D.R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer, New York, NY.

Cameron, A.C. and Trivedi, P.K. (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press, Cambridge.

Cepeda Carrión, G., Henseler, J., Ringle, C.M. and Roldán, J.L. (2016), "Prediction-oriented modeling in business research by means of PLS path modeling", *Journal of Business Research*, Vol. 69 No. 10, pp. 4545-4551.

Chica, M. and Rand, W. (2017), "Building agent-based decision support systems for word-of-mouth programs: a freemium application", *Journal of Marketing Research*, Vol. 54 No. 5, pp. 752-767.

Chin, W.W. (1998), "The partial least squares approach to structural equation modeling", in Marcoulides, G.A. (Ed.), *Modern Methods for Business Research*, Erlbaum, Mahwah, pp. 295-358.

Chin, W.W. (2010), "How to write up and report PLS analyses", in Esposito Vinzi, V., Chin, W.W., Henseler, J. and Wang, H. (Eds), *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Vol. 2, Springer Handbooks of Computational Statistics Series. Springer, Heidelberg, pp. 655-690.

Danks, N. and Ray, S. (2018), "Predictions from partial least squares models", in Ali, F., Rasoolimanesh, S.M. and Cobanoglu, C. (Eds), *Applying Partial Least Squares in Tourism and Hospitality Research*, Emerald Publishing Limited, Bingley, pp. 35-52.

Danks, N., Ray, S. and Shmueli, G. (2017), "Evaluating the predictive performance of composites in PLS path modeling", Working Paper, available at: SSRN: https://ssrn.com/abstract=3055222

Diamantopoulos, A. and Siguaw, J.A. (2000), *Introducing LISREL*, Sage, Thousand Oaks, CA.

Diamantopoulos, A. and Siguaw, J.A. (2006), "Formative vs reflective indicators in measure development: does the choice of indicators matter?", *British Journal of Management*, Vol. 13 No. 4, pp. 263-282.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P. and Kaiser, S. (2012), "Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective", *Journal of the Academy of Marketing Science*, Vol. 40 No. 3, pp. 434-449.

Evermann, J. and Tate, M. (2016), "Assessing the predictive performance of structural equation model estimators", *Journal of Business Research*, Vol. 69 No. 10, pp. 4565-4582.

Felipe, C., Roldán, J.L. and Leal-Rodríguez, A. (2017), "Impact of organizational culture values on organizational agility", *Sustainability*, Vol. 9 No. 12, p. 2354.

Franke, G. and Sarstedt, M. (2019), "Heuristics versus statistics in discriminant validity testing: a comparison of four procedures", *Internet Research*, forthcoming.

Fuchs, C. and Diamantopoulos, A. (2009), "Using single-item measures for construct measurement in management research. Conceptual issues and application guidelines", *Die Betriebswirtschaft*, Vol. 69 No. 2, pp. 197-212.

Geisser, S. (1974), "A predictive approach to the random effects model", *Biometrika*, Vol. 61 No. 1, pp. 101-107.

Geweke, J. and Meese, R. (1981a), "Estimating regression models of finite but unknown order", *Journal of Econometrics*, Vol. 16 No. 1, pp. 162.

Geweke, J. and Meese, R. (1981b), "Estimating regression models of finite but unknown order", *International Economic Review*, Vol. 22 No. 1, pp. 55-70.

Gregor, S. (2006), "The nature of theory in information systems", *MIS Quarterly*, Vol. 30 No. 3, pp. 611-642.

Hair, J.F., Black, W.C., Babin, B. and Anderson, R. (2018), *Multivariate Data Analysis*, Cengage, London, UK.

Hair, J.F., Hollingsworth, C.L., Randolph, A.B. and Chong, A.Y.L. (2017a), "An updated and expanded assessment of PLS-SEM in information systems research", *Industrial Management and Data Systems*, Vol. 117 No. 3, pp. 442-458.

Hair, J.F., Hult, G.T.M., Ringle, C.M. and Sarstedt, M. (2017b), *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 2nd ed., Sage, Thousand Oaks, CA.

Hair, J.F., Sarstedt, M., Pieper, T.M. and Ringle, C.M. (2012a), "The use of partial least squares structural equation modeling in strategic management research: a review of past practices

and recommendations for future applications", *Long Range Planning*, Vol. 45 Nos 5/6, pp. 320-340.

Hair, J.F., Sarstedt, M., Ringle, C.M. and Mena, J.A. (2012b), "An assessment of the use of partial least squares structural equation modeling in marketing research", *Journal of the Academy of Marketing Science*, Vol. 40 No. 3, pp. 414-433.

Henseler, J., Ringle, C.M. and Sarstedt, M. (2015), "A new criterion for assessing discriminant validity in variance-based structural equation modeling", *Journal of the Academy of Marketing Science*, Vol. 43 No. 1, pp. 115-135.

Hofman, J.M., Sharma, A. and Watts, D.J. (2017), "Prediction and explanation in social systems", *Science*, Vol. 355 No. 6324, pp. 486-488.

Homburg, C., Vomberg, A., Enke, M. and Grimm, P.H. (2015), "The loss of the marketing department's influence: is it happening? And why worry?", *Journal of the Academy of Marketing Science*, Vol. 45 No. 1, pp. 1-13.

Jöreskog, K.G. (1978), "Structural analysis of covariance and correlation matrices", *Psychometrika*, Vol. 43 No. 4, pp. 443-477.

Jöreskog, K.G. and Wold, H.O.A. (1982), "The ML and PLS techniques for modeling with latent variables: historical and comparative aspects", in Wold, H.O.A. and Jöreskog, K.G. (Eds), *Systems under Indirect Observation, Part I*, North-Holland, Amsterdam, pp. 263-270.

Kim, M.-J., Chung, N. and Lee, C.-K. (2011), "The effect of perceived trust on electronic commerce: shopping online for tourism products and services in South Korea", *Tourism Management*, Vol. 32 No. 2, pp. 256-265.

Kock, N. and Hadaya, P. (2018), "Minimum sample size estimation in PLS-SEM: the inverse square root and gamma-exponential methods", *Information Systems Journal*, Vol. 28 No. 1, pp. 227-261.

Lehmann, D.R., McAlister, L. and Staelin, R. (2011), "Sophistication in research in marketing", *Journal of Marketing*, Vol. 75 No. 4, pp. 155-165.

Lohmöller, J.-B. (1989), *Latent Variable Path Modeling with Partial Least Squares*, Physica, Heidelberg.

McQuarrie, A.D.R. and Tsai, C.-L. (1998), *Regression and Time Series Model Selection*, World Scientific, Singapore.

Makridakis, S. (1993), "Accuracy measures: theoretical and practical concerns", *International Journal of Forecasting*, Vol. 9 No. 4, pp. 527-529.

Mooi, E.A., Sarstedt, M. and Mooi-Reci, I. (2018), *Market Research: The Process, Data, and Methods Using Stata*, Springer, Heidelberg.

Musa, H.D., Yacob, M.R., Abdullah, A.M. and Ishak, M.Y. (2018), "Enhancing subjective well-being through strategic urban planning: development and application of community happiness index", *Sustainable Cities and Society*, Vol. 38, pp. 184-194.

Myung, I.J. (2000), "The importance of complexity in model selection", *Journal of Mathematical Psychology*, Vol. 44 No. 1, pp. 190-204.

Nau, R. (2016), "Statistical forecasting: notes on regression and time series analysis", available at: https://people.duke.edu/~rnau/compare.htm (accessed 18 February 2019).

Oliveira, P.S., da Silva, L.F., d., ¨Silva, D., Tecilla, M.C. and da Silva, R.C. (2018), "World class manufacturing operations management: scale development and LHEMI model proposition", *International Journal of Innovation and Technology Management*, Vol. 15 No. 05, pp. 5, 1850042.

Pontius, R.G., Thontteh, O. and Chen, H. (2008), "Components of information for multiple resolution comparison between maps that share a real variable", *Environmental and Ecological Statistics*, Vol. 15 No. 2, pp. 111-142.

Rasoolimanesh, S.M. and Ali, F. (2018), "Editorial: partial least squares (PLS) in hospitality and tourism research", *Journal of Hospitality and Tourism Technology*, Vol. 9 No. 3, pp. 238-248.

Reibstein, D.J., Day, G. and Wind, J. (2009), "Guest editorial: is marketing academia losing its way?", *Journal of Marketing*, Vol. 73 No. 4, pp. 1-3.

Rigdon, E.E. (1998), "Structural equation modeling", in Marcoulides, G.A. (Ed.), *Modern Methods for Business Research*, Erlbaum, Mahwah, pp. 251-294.

Rigdon, E.E. (2012), "Rethinking partial least squares path modeling: in praise of simple methods", *Long Range Planning*, Vol. 45 Nos 5/6, pp. 341-358.

Ringle, C.M., Wende, S. and Becker, J.-M. (2015), *SmartPLS 3*, SmartPLS, Bönningstedt.

Ringle, C.M., Sarstedt, M., Mitchell, R. and Gudergan, S.P. (2019), "Partial least squares structural equation modeling in HRM research", *The International Journal of Human Resource Management*, Forthcoming.

Saayman, M., Li, G., Uysal, M. and Song, H. (2018), "Tourist satisfaction and subjective well-being: an index approach", *International Journal of Tourism Research*, Vol. 20 No. 3, pp. 388-399.

Salzberger, T., Sarstedt, M. and Diamantopoulos, A. (2016), "Measurement in the social sciences: where C-OAR-SE delivers and where it does not", *European Journal of Marketing*, Vol. 50 No. 11, pp. 1942-1952.

Sarstedt, M. and Mooi, E.A. (2019), *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*, Springer, Heidelberg.

Sarstedt, M., Ringle, C.M. and Hair, J.F. (2017), "Partial least squares structural equation modeling", in Homburg, C., Klarmann, M. and Vomberg, A. (Eds), *Handbook of Market Research*, Springer, Heidelberg.

Sarstedt, M., Diamantopoulos, A., Salzberger, T. and Baumgartner, P. (2016a), "Selecting single items to measure doubly-concrete constructs: a cautionary tale", *Journal of Business Research*, Vol. 69 No. 8, pp. 3159-3167.

Sarstedt, M., Ringle, C.M., Henseler, J. and Hair, J.F. (2014), "On the emancipation of PLS-SEM: a commentary on Rigdon (2012)", *Long Range Planning*, Vol. 47 No. 3, pp. 154-160.

Sarstedt, M., Hair, J.F., Ringle, C.M., Thiele, K.O. and Gudergan, S.P. (2016b), "Estimation issues with PLS and CBSEM: where the bias lies!", *Journal of Business Research*, Vol. 69 No. 10, pp. 3998-4010.

Sarstedt, M., Ringle, C.M., Cheah, J.-H., Ting, H., Moisescu, O.I. and Radomir, L. (2019), "Structural model robustness checks in PLS-SEM", *Tourism Economics*, forthcoming.

Schwarz, G. (1978), "Estimating the dimensions of a model", *The Annals of Statistics*, Vol. 6 No. 2, pp. 461-464.

Sethi, V. and King, W.R. (1991), "Construct measurement in information systems research: an illustration in strategic systems", *Decision Sciences*, Vol. 22 No. 3, pp. 455-472.

Sharma, P.N., Shmueli, G., Sarstedt, M., Danks, N. and Ray, S. (2019a), "Prediction-oriented model selection in partial least squares path modeling", *Decision Sciences*, forthcoming.

Sharma, P.N., Shmueli, G., Sarstedt, M., Kim, K.H. and Thiele, K.O. (2019b), "PLS-based model selection: the role of alternative explanations in MIS research", *Journal of the Association for Information Systems*, Vol. 20 No. 3, pp. 346-397.

Shmueli, G. (2010), "To explain or to predict?", *Statistical Science*, Vol. 25 No. 3, pp. 289-310.

Shmueli, G. and Koppius, O.R. (2011), "Predictive analytics in information systems research", *MIS Quarterly*, Vol. 35 No. 3, pp. 553-572.

Shmueli, G., Ray, S., Velasquez Estrada, J.M. and Chatla, S.B. (2016), "The elephant in the room: evaluating the predictive performance of PLS models", *Journal of Business Research*, Vol. 69 No. 10, pp. 4552-4564.

Steenkamp, J.B.E.M. and Baumgartner, H. (2000), "On the use of structural equation models for marketing modeling", *International Journal of Research in Marketing*, Vol. 17 Nos 2/3, pp. 195-202.

Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions", *Journal of the Royal Statistical Society*, Vol. 36 No. 2, pp. 111-147.

Streukens, S. and Leroi-Werelds, S. (2016), "Bootstrapping and PLS-SEM: a step-by-step guide to get more out of your bootstrap results", *European Management Journal*, Vol. 34 No. 6, pp. 618-632.

Svensson, G., Ferro, C., Høgevold, N., Padin, C., Varela, J.C.S. and Sarstedt, M. (2018), "Framing the triple bottom line approach: direct and mediation effects between economic, social and environmental elements", *Journal of Cleaner Production*, Vol. 197 No. 1, pp. 972-991.

Teo, T.S.H. and Yeong, Y.D. (2003), "Assessing the consumer decision process in the digital marketplace", *Omega*, Vol. 31 No. 5, pp. 349-363.

Tofallis, C. (2015), "A better measure of relative prediction accuracy for model selection and model estimation", *Journal of the Operational Research Society*, Vol. 66 No. 8, pp. 1352-1362.

Vanwinckelen, G. and Blockeel, H. (2012), "On estimating model accuracy with repeated cross-validation", in *21st Belgian-Dutch Conference on Machine Learning*, *Ghent*, pp. 39-44.

Willmott, C.J. and Matsuura, K. (2005), "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance", *Climate Research*, Vol. 30 No. 1, pp. 79-82.

Witten, I., Frank, E., Hall, M. and Pal, C.J. (2016), *Data Mining: Practical Machine Learning Tools and Technique*, Morgan Kaufmann, Burlington, MA.

Wold, H.O.A. (1985), "Partial least squares", in Kotz, S. and Johnson, N.L. (Eds), *Encyclopedia of Statistical Sciences*, Wiley, New York, NY, pp. 581-591.

Appendix

Measurement items

(1) *Attitude (ATT)*:
- Online journey shopping is a good idea (*att*)

(2) *Communicability (COMM)*:
- I have heard about people booking journeys online many times. (*comm1*)
- Many friends have purchased journeys online. (*comm2*)
- It is common for people to purchase journeys online. (*comm3*)

(3) *Convenience (CONV)*:
- Purchasing journeys online makes me less dependent of opening hours. (*conv1*)
- Purchasing journeys online has easy payment procedures. (*conv2*)
- Purchasing journeys online is more convenient than regular shopping, as I can do it anytime and anywhere. (*con3*)

(4) *Time saving (TS)*:
- Purchasing journeys online enables me to complete my shopping quickly. (*ts1*)
- I can save time by purchasing journeys online. (*ts2*)
- Purchasing travel online takes less time than purchasing them at travel agencies. (*ts3*)

(5) *Financial advantages (FA)*:
- I save money by purchasing journeys online. (*fa1*)
- Online journey shopping provides more discounts than offline travel purchasing. (*fa2*)
- Generally, travel websites offer tourism products at cheaper prices. (*fa3*)

(6) *Produce variety (PV)*:
- There is a larger choice of travel products available when purchasing online. (*pv1*)
- The Internet allows me to purchase travel services that are not available offline. (*pv2*)
- I can design a custom-made trip by purchasing a journey online. (*pv3*)

(7) *Enjoyment (ENJOY)*:
- Purchasing journeys online is more exciting than purchasing them offline. (*enjoy1*)
- I enjoy purchasing journeys online more than purchasing them offline. (*enjoy2*)

(8) *Trust (TRUST)* (Kim *et al.*, 2011):
- The chance of having a technical failure during an online transaction is quite small. (*trust1*)
- I believe most e-commerce travel websites will do their outmost to benefit customers. (*trust2*)
- I believe online travel sites are trustworthy. (*trust3*)
- Internet shopping is reliable. (*trust4*)
- Internet shopping can be trusted and there are no uncertainties. (*trust5*)

(9) *Intention to purchase travel online (IPTO)* (adapted from Bigné *et al.*, 2010; Teo and Yeong, 2003):
- If you were to purchase journeys the probability that you will purchase them online is high. (*ipto1*)
- I expect to purchase journeys online in the near future (*ipto2*)

Demographic variable	Category	Frequency	(%)
Gender	Male	143	40.9
	Female	207	59.1
Ethnicity	Malay	221	63.1
	Chinese	86	24.6
	Indian	20	5.7
	Others	23	6.6
Age	21-30 years old	192	54.9
	31-40 years old	120	34.3
	41-50 years old	27	7.7
	51-60 years old	11	3.1
Level of income	RM2,001-RM3,000	153	43.7
	RM3,001-RM4,000	69	19.7
	RM4,001-RM5,000	31	8.9
	RM5,001-RM6,000	32	9.1
	RM6,001-RM7,000	22	6.3
	RM7,001 and above	43	12.3
How frequently did you purchase a journey online?	Yearly	230	65.7
	Half-yearly	68	19.4
	Quarterly	35	10.0
	Monthly	17	4.9
Purpose of purchasing a journey online	Family-trip	233	66.6
	Business-trip	33	9.4
	Honeymoon-trip	18	5.1
	Friends-trip	66	18.9

Table AI. Sample demographics

Corresponding author
Marko Sarstedt can be contacted at: marko.sarstedt@ovgu.de