1 A Multi-trait multi-method validity test of partworth estimates

Wagner Kamakura and Muammer Ozer

1.1 Introduction

Conjoint analysis has already been widely accepted by marketing researchers as a popular instrument for the measurement of consumer preferences. Typical applications of conjoint analysis include new product design based on the relationship between product features and predicted choice behavior, benefit segmentation based on attribute preferences, etc. The popularity of conjoint analysis among marketing researchers hinges on the belief that it produces valid measurements of consumer preferences for the features of a product or service, and that it provides accurate predictions of choice behavior.

Given this importance, a vast literature has already emerged on the validity of conjoint analysis for the measurement of preferences. However, as we will discuss it in more details in our literature review section, these tests usually involve collecting data from two separate conjoint tasks (usually in the same interview), estimating part-worths based on the first task, using the estimates to make predictions about the second task, and measuring predictive fit. We argue that these tests are more akin to test-retest reliability assessments than validation tests. Our literature review also indicates that comparative conjoint studies have yielded inconclusive results partly due to the different validity measures used as a basis for validation.

The purpose of this chapter is to compare various conjoint models across their part-worth estimates based on actual behavior. Because we are comparing those methods across their part-worth estimates, we use a Multitrait-Multimethod (MTMM) framework to assess the
relationships among the methods and the part-worth estimates. We test the relationships by using both the traditional MTMM analysis and a direct product methodology. The following section presents our literature review. After that, we provide a summary of the MTMM and direct product methodologies. Then, we present details about our research design and the conjoint models that we used to generate part-worth estimates. Finally, we discuss our results and conclude the chapter with managerial and research implications.

1.2 Literature Review

Previous studies have proposed a number of methods to improve the validity and reliability of conjoint analysis. For example, Hagerty (1985) used a Q-type factor analysis to determine optimum weights that optimize the expected mean squared error of prediction in a validation sample. He tested his methodology by using both synthetic and real data. He manipulated the amount of overlaps in the clusters of part-worth estimates in two Monte Carlo simulations. He also asked student subjects to rank-order eighteen job descriptions based on five attributes. Two of the eighteen job descriptions were used as a holdout sample. He compared his results with those of non-overlapping clustering and individual-level clustering. Based on Mean Squared Error (MSE) and the first choice prediction criteria, he showed that his methodology yielded higher predictive accuracy.

In another study, Kamakura (1988) used an agglomerative hierarchical method for simultaneous segmentation and estimation of conjoint models. He used a least squares procedure to identify segments that maximize the predictive validity of the segment-level part-worth estimates. He compared his results with those of the two-stage segmentation procedure (individual-level estimation and clustering of subjects based on part-worth estimates) by using both synthetic and empirical data based on holdout samples. His synthetic data included simulated preference rankings whereas his real data consisted of preference ratings of twenty-seven full-profile descriptions of checking account services. The holdout sample included eight profiles with the same attributes. He concluded that his methodology and the two-stage segmentation procedure yielded similar part-worth estimates, but his
proposed methodology gave consistently more accurate results. Ogawa (1987) also suggested a similar procedure, but used a logit-based estimation methodology. By using both simulated data and a set of preference data for Japanese automobiles, he was able to show that his part-worth estimates were internally valid. Similar to earlier studies, his validity assessment was also based on a holdout sample.

Wedel and Steenkamp (1989) proposed a fuzzy clusterwise regression algorithm to allow consumers to possess partial membership in multiple segments. Wedel and Steenkamp (1989) compared their procedure with a clusterwise regression and Hagerthy's optimal weighting method. They first used a simulated data to validate the computational efficiency of their method. They later used a data set about customer satisfaction with respect to eight stock market scenarios and another data set for meat products. The results of the simulation showed that the methodology was computationally sound. In addition, judging by the percentage of first choices accurately predicted in a holdout sample with similar tasks, they concluded that their method and the clusterwise regression gave consistent results. They also concluded that their results were slightly better than those of Hagerthy's method for well-defined clusters, but were worse for diffuse clusters. Similarly, DeSarbo et al. (1992) introduced a latent class methodology that allowed overlapping clusters for simultaneous segmentation and parameter estimation by using a mixture of multivariate conditional normal distributions. They successfully applied the methodology to a conjoint experiment on remote controls for cars, simultaneously identifying segments and generating part-worth estimates within each segment. When they compared their methodology with various OLS procedures, they found that their latent class procedure outperformed them based on a likelihood-ratio test and a goodness of fit index (R-square).

As an example of incorporating exogenous variables into the segmentation process, Kamakura, Wedel and Agrawal (1994) presented a multinomial mixture model for the external analysis of rank-order, pick-any and conjoint choice data. They used their model to simultaneously determine market segments based on consumer characteristics and to generate part-worth estimates for each segment. They used synthetic data to assess the performance of their methodology in recovering "true" parameters in a sample. They also used a conjoint experiment about banking services to assess the predictive validity based on a holdout sample. Both the estimation and the holdout samples included nine profiles that were
equivalent, but distinct. They compared their model with a naive model, an aggregate rank order logit model with dummy variables, and a latent class rank order logit model without concomitant variables. The results provided support for the internal and predictive validity of the proposed methodology.

Despite the proven performance of the models, the results of the comparative studies have been inconclusive. For instance, Hagerty (1985) and Kamakura (1988) found that their proposed methodologies outperformed the alternative OLS methodology. However, Green and Helsen (1989) conducted a conjoint experiment about student apartments and concluded that neither Hagerty's optimal weighting methodology nor Kamakura's method led to higher predictive validities than were obtained by conventional OLS. Their validation was based on a holdout data that was collected during the same experiment. The estimation set included eighteen full profile descriptions whereas the holdout sample had sixteen profiles with the same attributes. Similarly, Green, Krieger and Schaffer (1993) used three different studies to compare the predictive accuracy of Hagerty's optimal weighting methodology with those from individual OLS estimation. Based on a holdout validation sample, they showed that the optimal weighting methodology did not outperform the OLS method.

In an extensive Monte Carlo study, Vriens, Wedel and Wilms (1996) manipulated the number of subjects, number of profiles, number of segments, error variance, segment homogeneity, and segment similarity to compare nine methods of metric conjoint segmentation based on parameter recovery, goodness of fit and predictive power. The methods included the traditional two-stage approaches with (TTSWA) and without (TTSKM) hierarchical clustering, the alternative two-stage methods using Ward's clustering (ATSWA) and a K-means clustering procedure (ATSKM), the optimal weighting methods (OW), the optimal weighting method followed by a K-means clustering procedure (OWKM), the clusterwise regression procedure (CR), the fuzzy clusterwise regression procedure (FCR), and the latent class normal distribution model (LCN). In terms of goodness of fit, the results indicated that the two-stage clustering procedures (TTSWA, TTSKM, ATSWA, and ATSKM) outperformed OWKM, FR, and LCN. However, with respect to parameter recovery, methods that integrate segmentation and estimation aspects of conjoint analysis (integrated conjoint segmentation methods) outperformed the two-stage clustering procedures, most notably, the latent class methodology of DeSarbo et al. (1992) performed best among the integrated methods. Finally, the tested
methods had similar predictive power based on an assessment with a holdout sample.

1.2.1 Validation of Part-Worth Estimates

Among the important managerial applications of conjoint analysis are the design of new products, development of advertising and marketing strategies, and the identification of relevant market segments for product targeting and positioning. Managers achieve these objectives by considering the part-worth estimates obtained from a conjoint study. Despite the importance of part-worth estimates in marketing, most research studies have investigated the ability of conjoint models in predicting overall preferences for bundles of attributes, as opposed to investigating the validity of part-worth estimates. Previous researchers such as Vriens, Wedel, and Wilms (1996) looked at parameter recovery measured by the root-mean-squared-error between the "true" (simulated) and estimated values of the part-worths. Our study aims to compare the part-worth estimates obtained from different conjoint models with actual consumer behavior that is directly related to these estimates.

There are a few studies that have already looked into the validation of part-worth estimates. However, they have primarily focused on the consistency of the part-worth estimates across estimation procedures. For example, in an earlier study about consumer preferences for banks in which to open a checking account, Jain et al. (1979) compared the part-worth estimates obtained from different conjoint estimation techniques such as MONANOVA, JOHNSON, LINMAP, LOGIT, and OLS. Based on a holdout sample of conjoint profiles, they showed that the methods yielded significantly different part-worth estimates across different data collection methods. In addition, they showed that the LINMAP procedure was effective in predicting the first choice in the holdout sample whereas the OLS procedure was more effective in predicting the least preferred choice in the sample. Leigh et al. (1984) also compared the part-worth coefficients estimated by using different procedures such as rank-order, paired comparison, graded paired comparison, and ranking scale with the weights elicited through a self-explicated procedure for hand-held calculators. The results failed to provide support for the presumed greater reliability and validity of the tested procedures over the self-explicated procedure. The reliability was assessed by test-retest comparison whereas
the predictive validity was measured based on a simulated choice (raffle) in a separate interview.

In a conjoint experiment involving student apartments, Akaah (1991) compared the predictive performance of self-explicated, traditional conjoint and hybrid conjoint models under alternative data collection modes including in-person interviews, mail questionnaires and telephone interviews. By using eighteen estimation profiles and six holdout profiles, they showed that the self-explicated and traditional conjoint models gave fairly similar attribute importance weights across the different data collection modes. In addition, Darmon and Rouzies (1991) conducted a conjoint simulation to investigate the internal validity of part worth estimates across different design (full vs. fractional) and different estimation procedures (LINMAP, MONANOVA and OLS). By comparing the part-worth estimates generated by known functional forms (“true” values) with those generated by a conjoint analysis procedure, they showed that the part-worth estimates were more valid with full design rather than with a fractional factorial design, and with OLS rather than other estimation procedure when a fractional design was mandatory. In a subsequent simulation study, Darmon and Rouzies (1994) also investigated the role of error in the reliability and internal validity of part worth estimates. The results indicated that when there is a low level of error in the input data less important attributes are underestimated whereas when there is a high level of error in the input data less important attributes are overestimated.

1.2.2 Validation Measures

Researchers argued that the inconclusive results of previous comparative studies could be due to several factors. Among the most important factors affecting the results is the use of one or two data sets and/or the use of a small holdout sample (Hagerty 1993, Vriens, Wedel and Wilms 1996). For example, Hagerty (1993) stated that the type of brands included in a holdout sample could affect the validity results and urged researchers to consider multiple holdout samples for cross validation. It can also be due to various validation measures used in the literature.

Previous studies have used different validity measures. Given the difficulties in externally validating conjoint experiments, many studies have emphasized internal validity which can be measured by a test/re-test
reliability analysis and/or cross-validation, i.e., the ability of a model to predict the rankings or the first choice in a hold-out sample (Green, Krieger, and Agarwal 1993). Alternatively, some studies also used Monte Carlo simulations to validate or compare alternative conjoint methodologies (Vriens, Wedel and Wilms 1996) on synthetic data.

In this chapter, we attempt to assess the external and convergent validity of the preference measurements obtained from various conjoint methods. We used actual behavior of individuals for comparison. More specifically, we compared the preferences of customers of a major bank for a factorial design of account characteristics with their actual banking behavior on those dimensions that were measured immediately prior to the conjoint study and were totally independent from the conjoint task. We believe that using actual behavior for validating and/or comparing alternative conjoint methods can reduce some of the concerns associated with holdout samples and other alternative validation measures.

There are a few studies that have already used actual behavior for validation. For example, Krishnamurthi (1988) asked MBA students and their spouses to rank-order 28 hypothetical job descriptions. Three months later, he asked the couples to rank-order the job offers that they received. The predictive validity was calculated based on the actual job choices that the models predicted. Similarly, Srinivasan and Park (1997) conducted a conjoint study to identify how important different factors were for MBA students in choosing among job offers. Three months later they asked the participants about the number of job offers they had and the one they selected. The predictive validity was assessed based on a comparison between the actual job choices and the predictions of the conjoint model. Although this can also be a useful measure, the three-month time lag can involve changes in the environment and in individuals’ preferences.

Some researchers have also used raffles to simulate an actual choice environment. For example, Leigh et al. (1984) first conducted a conjoint experiment for hand-held calculators. Two weeks later, their subjects participated in a raffle for a calculator of their choice from a predetermined set of ten calculators. The subjects’ choice in the raffle represented actual behavior and was a base for predictive validity. Although these measures provide a powerful way of assessing the validity of conjoint studies, the results can be subject to a carry-over bias. In other words, once the respondents participate in a study, they tend to remember their answers and be consistent if they are required to participate in a similar study.
(Morwitz et al. 1993). This bias can be reduced by having a longer time period between the two studies. However, as the time gap gets larger, the conditions and individuals’ preferences can change. Thus, we will not know whether the difference between the estimation and actual behavior is due to the method used or due the changes in the conditions. Furthermore, these studies focused on predictive validity of the composite utilities, rather than on the measurement of preferences for each attribute.

1.3 The Multitrait Multimethod (MTMM) Methodology

Because we are interested in comparing the validity of various conjoint methodologies in measuring preferences for various attributes, we compared the methods within an MTMM framework. The MTMM analysis was first introduced by Campbell and Fiske (1959) for construct validation. A typical MTMM matrix includes correlations among multiple traits (concepts) measured by multiple methods and enables researchers to determine the extent of similarities of the methods (convergence) and the extent of uniqueness of the traits (discrimination).

Campbell and Fiske (1959) also suggested specific criteria for convergent and discriminant validity in analyzing an MTMM matrix. Convergent validity is achieved when the correlations between attempts to measure the same concept with different methods (i.e., montrait-heteromethod correlations) are significantly different from zero and sufficiently large. On the other hand, discriminant validity is achieved when:

1. the correlations between attempts to measure the same concept with different methods (i.e., the montrait-heteromethod correlations) are larger than the correlations between attempts to measure different concepts with different methods (i.e., the heterotrait-heteromethod correlations);

2. the correlations between attempts to measure the same concept with different methods (i.e., the montrait-heteromethod correlations) are larger than the correlations between different concepts measured by the same method (i.e., the heterotrait-monomethod correlations); and finally
3 the patterns of correlations between different concepts are consistent under the same or different methods (Bagozzi and Yi 1991).

Despite providing these criteria, Campbell and Fiske (1959) did not specify any objective mechanism to test them. The criteria are subjectively assessed by researchers, thus leading to ambiguous interpretations. Researchers later developed statistical procedure to analyze MTMM matrices, most notably Confirmatory Factor Analysis (CFA) and direct product decomposition. CFA decomposes the total variation into (1) the variation due to differences in individual trait scores, (2) the variation due to differences in systematic biases induced by methods used and (3) the variation due to random error (Bagozzi and Yi 1991). The CFA methodology assumes that the total variation is a linear combination (additive) of the variations due to traits, methods and error. Researchers have questioned the validity of this assumption by showing some level of interaction between traits and methods (Campbell and O’Connell 1967; Lastovicka et al. 1990; Kumar and Dillon 1992).

As an alternative, Browne (1984) proposed a methodology called Direct Product Model for analyzing MTMM matrices based on a multiplicative decomposition of the trait and method effects. Under this methodology, the observed covariance matrix is expressed as

$$\Sigma = Z(P_M \otimes P_T + E^2)Z$$

Where $\Sigma$ is the population covariance matrix for the observed scores; $P_M$ and $P_T$ are nonnegative definite matrices of method and trait correlations, respectively; $E^2$ is a nonnegative definite diagonal matrix for the unique variances; $Z$ is a nonnegative definite diagonal matrix of scale constant; and $\otimes$ represents a right direct (Kronecker) product. Note that the elements of $P_M$ and $P_T$ represent multiplicative components of common score correlations, that is correlations corrected for attenuation as opposed to observed score correlations. In addition, the identification of scale factor estimates requires one equality constraint per method. For instance, one may select a trait and set all its scale parameters (corresponding diagonal elements of $Z$) equal to unity (Wothke and Browne 1990).

An important aspect of the direct product model is that the $P_M$ and $P_T$ matrices can be directly used to test the convergent and discriminant
validity of constructs as suggested by Campbell and Fiske (1959). More specifically, convergent validity is achieved when the correlations among methods in \( P_M \) are positive and large. On the other hand, discriminant validity is achieved when (1) the correlations among traits in \( P_T \) are low; (2) the method correlations in \( P_M \) are greater than the trait correlations in \( P_T \); and (3) the direct product model fits the data (Bagozzi and Yi 1991).

Previous studies have compared the additive and multiplicative approaches to analyzing MTMM matrices (Lastovicka et al. 1990; Bagozzi and Yi 1991; Kumar and Dillon 1992). Despite the inconclusive results, the studies suggested that the use of a particular methodology should be based on theory. Furthermore, Kumar and Dillon (1992) argued that when the primary purpose is to assess the extent to which the MTMM data satisfy the Campbell and Fiske (1959) criteria, the direct product model should suffice. As we will discuss it further in the following section in more details, we compare the part-worth estimates of various conjoint models for banking services. In addition, we also compare the part-worth estimates with actual behavior. Because we are using actual behavior as one of the methods the part-worth estimates interact with the traits. In other words, the part-worth estimates of the conjoint models should be different than actual behavior. Furthermore, because our primary focus is to test the criteria set by Campbell and Fiske (1959) we use a direct product model to conduct the MTMM analysis.

The following section briefly explains the research methodology and the various conjoint models used in the MTMM analysis. After that, we present the results of the MTMM analysis.

1.4 Description of Data and Methods

Our Multitrait Multi-Method analysis will be based on a comparison of part-worth estimates for multiple attributes obtained from four different estimation methods, with actual behavior, directly related to the attributes preferences measured in the conjoint task, observed from the same consumers. For this purpose, we use the same data utilized by Kamakura et al. (1994) to illustrate their latent-class conjoint segmentation model. This
commercial study involved four attributes for checking accounts, each manipulated in three levels:

- **MINBAL**: minimum balance required to exempt the customer from a monthly service fee ($0, $500 or $1,000)
- **CHECK**: cost to the customer per checked issued (0c, 15c or 35c)
- **FEE**: monthly service fee ($0, $3 or $6)
- **ATM**: availability and cost of ATM (no access, free ATM, or paid ATM @ 75c per transaction)

Two distinct but equivalent sets of 9 hypothetical checking accounts were generated using a fractional factorial design. A random sample of 269 customers from the bank was asked to rank the nine accounts from the first set in order of preference. After a series of other questions related to the same commercial study, the respondents were asked to rank the second set of profiles. These two similar tasks provided us the opportunity of replicating the test-retest reliability comparisons that are commonly reported in the conjoint literature.

In addition to the two conjoint data sets, we also obtained the following information (in disguised and anonymous format) from the bank, regarding the participants’ banking behavior in the 6 months immediately prior to the conjoint task:

- **BALANCE**: average balance kept in the account (earning 5% interest)
- **NCHECK**: number of checks issued per month
- **SRVFEES**: number of times the customer paid a monthly service fee
- **NATM**: number of ATM transactions per month

This combination of conjoint and behavior data provides us with a rare opportunity of testing for the external validity of the preference measurements obtained from conjoint analysis using a variety of estimation methods. Preferences for the levels of the four attributes in the conjoint task can be viewed as multiple traits. On the other hand, measurements for each of these attributes obtained from different estimation methods, along with the actual banking behavior associated with these attributes represent multiple measurement methods, leading to the classic Multitrait Multi-methods approach for assessing the validity of measurement instruments. Given that we know the actual behavior of the respondents, we already have an idea of the inter-trait correlations. For
instance, people who are sensitive to minimum balance should be more willing to pay a monthly fee whereas people who do not care about minimum balance (i.e., people who have money in their accounts) do not want to pay a monthly fee. In addition, given the inconclusive results of the previous comparative conjoint validation studies, we are more interested in the inter-method correlations. Moreover, one of our “methods” is the actual banking behavior observed immediately prior to the conjoint task, providing a benchmark for assessing predictive validity.
Figure 1  Least squares regressions (OLS) to the data from each of the 269 subjects.

The following measurement methods are used in our MTMM analysis:

1. individual OLS estimates,
2. Kamakura’s (1988) agglomerative hierarchical regression (KAM),
3. Hagerty’s (1985) Q-factor methodology (HAG), and
4. Latent Class Rank Logit model (Kamakura et al. 1994).
The first method (OLS) involves the estimation of a linear regression for each respondent across all nine profiles, using the (inverted) preference ranking as the dependent variable. The next two models (KAM and HAG) attempt to improve the predictive validity of the part-worth estimates by estimating them within homogeneous groups of consumers (KAM), or by obtaining an optimal partitioning of the sample via Q-factor analysis. The emphasis in these two techniques is obtaining individual-level estimates that would maximize the ability to predict preferences. The main purpose of the latent-class rank logit model, on the other hand, is market segmentation, i.e., identify relatively groups of consumers who are relatively homogeneous in their preferences for the attributes.

Application of least squares regressions (OLS) to the data from each of the 269 subjects leads to the part-worth estimates displayed in Figure 1. The partworths for MINBAL, CHECK and FEE are the regression coefficients for the (mean-centered) attributes, while the partworths for ATM are the coefficients for the effects-coded dummies for Free ATM and Paid ATM.

![Predictive Validity Index](image)

**Figure 2** Predictive validity indices.

The hierarchical clusterwise regression approach (KAM) joins consumers into hierarchical segments to maximize a predictive validity index, which indicates the ability of predicting individual-level preferences using cluster-level estimates of partworths. Application of the model to our data led to the predictive validity indices displayed in Figure 2. Based on this
figure, and on the purpose of maximizing predictive accuracy, we chose the 23-cluster solution, which is not appropriate for segmentation purposes, but produces the highest expected predictive accuracy. The distribution of part-worth estimates across the 269 consumers are shown in Figure 3.

![Histograms of part-worth estimates](image)

**Figure 3** Distribution of part-worth estimates across the 269 consumers.

The method proposed by Hagerty (1985) (HAG) amounts to a Q-factor analysis of the between-subjects covariance matrix of preference ratings. The eigenvalues obtained from applying Hagerty's Q-factor analysis to our data are displayed in Figure 4. Based on these values, we decided for a 3-
factor solution. The individual-level estimates of partworths are summarized in Figure 5.

![Scree plot for Hagerty's model](image)

**Figure 4**  
Eigenvalues obtained from applying Hagerty's Q-factor analysis.

Application of the latent-class rank logit model led to 4 latent classes, chosen on the basis of the ICOMP criterion (see Kamakura et al. 1994 for details on this analysis). The individual-level estimates, based on this 4-class solution are summarized in Figure 6. The reader should note that while the other approaches being compared specify a linear model relating the attributes to the observed preferences, the latent-class rank logit model is applied to preference rankings. Consequently, the relationship between the attributes and the observed preferences is non-linear, and the estimated partworths are not directly comparable to the ones obtained from the other approaches.
Aside from the estimates shown in Figures 1, 3, 5 and 6, we also obtained estimates for each of the models using the second set of conjoint profiles. The correlations between these two measurements for the same constructs and subjects using the same method in two separate tasks provide us with the measures of test-retest reliability for each construct (partworth) and method.

As mentioned earlier, we also had data on the actual banking behavior of each of the 269 respondents on variables that are directly related to the estimated partworths. Consumers who are highly sensitive to minimum balance (extreme negative partworths for MINBAL) would be expected to
have lower average balances (BALANCE). Therefore, the partworths estimates for MINBAL obtained from any method should be positively correlated to BALANCE.

Figure 6  The individual-level estimates based on this 4-class solution.

Similarly, those who are highly sensitive to cost/check (extreme negative partworths for CHECK) would be expected to issue more checks per month (NCHECK). In order to maintain a positive correlation between partworth estimates and observed behavior, we invert the sign of the
partworth estimates for CHECK, which should then be positively correlated with NCHECK.

<table>
<thead>
<tr>
<th>Checks issued per month</th>
<th>Average balance in past year</th>
<th>MR. BALANCE</th>
<th>COSTCHECK</th>
<th>MONTHLY FEE</th>
<th>ATM ACCESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>OLS</td>
<td>KAM</td>
<td>HAG</td>
<td>LCA</td>
<td>ACT</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------------</td>
<td>-------------</td>
<td>------------</td>
<td>-------------</td>
<td>------------</td>
</tr>
<tr>
<td>52</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>100</td>
</tr>
<tr>
<td>OLS estimate for min. balance</td>
<td>39</td>
<td>87</td>
<td>39</td>
<td>94</td>
<td>76</td>
</tr>
<tr>
<td>KAM estimate for min. balance</td>
<td>39</td>
<td>94</td>
<td>76</td>
<td>39</td>
<td>99</td>
</tr>
<tr>
<td>HAG estimate for min. balance</td>
<td>39</td>
<td>99</td>
<td>85</td>
<td>39</td>
<td>93</td>
</tr>
<tr>
<td>LCA estimate for min. balance</td>
<td>39</td>
<td>99</td>
<td>85</td>
<td>39</td>
<td>93</td>
</tr>
<tr>
<td>Number of times paid fee</td>
<td>-54</td>
<td>-35</td>
<td>-35</td>
<td>-35</td>
<td>-37</td>
</tr>
<tr>
<td>LCA estimate for monthly fee</td>
<td>-10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of ATM transactions</th>
<th>0</th>
<th>2</th>
<th>5</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>31</th>
<th>-15</th>
<th>-14</th>
<th>-16</th>
<th>-8</th>
<th>23</th>
<th>14</th>
<th>12</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS importance of ATM</td>
<td>10</td>
<td>26</td>
<td>28</td>
<td>19</td>
<td>14</td>
<td>10</td>
<td>-3</td>
<td>-1</td>
<td>9</td>
<td>-44</td>
<td>1</td>
<td>13</td>
<td>2</td>
<td>7</td>
<td>39</td>
<td>13</td>
</tr>
<tr>
<td>KAM importance of ATM</td>
<td>12</td>
<td>27</td>
<td>32</td>
<td>22</td>
<td>14</td>
<td>6</td>
<td>-4</td>
<td>-2</td>
<td>9</td>
<td>-43</td>
<td>-6</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>39</td>
<td>10</td>
</tr>
<tr>
<td>HAG importance of ATM</td>
<td>27</td>
<td>54</td>
<td>56</td>
<td>65</td>
<td>59</td>
<td>12</td>
<td>42</td>
<td>43</td>
<td>45</td>
<td>16</td>
<td>-21</td>
<td>-33</td>
<td>-32</td>
<td>-46</td>
<td>-20</td>
<td>16</td>
</tr>
<tr>
<td>LCA importance of ATM</td>
<td>-10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-13</td>
<td>-8</td>
<td>-46</td>
<td>-41</td>
<td>-43</td>
<td>-82</td>
<td>16</td>
<td>24</td>
<td>23</td>
<td>24</td>
<td>44</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 1. MTMM Matrix

Customers who already pay monthly service fees (SRVFEES) should show less resistance to monthly fees (FEE) than those who rarely pay monthly service fees. Therefore, the partworth estimates for FEE should be positively correlated with the observed behavior SRVFEES, irrespective of the estimation method.

One should also expect customers with a large number of ATM transactions (NATM) to value this service more than those who are light users or non-users of it. In order to measure respondents’ perceived value for ATM’s, we create a new measure (ATMV) by subtracting the partworth for paid ATM (75c) from the partworth for free ATM. This new measure of preference for ATM’s (ATMV) should be positively correlated with the number of ATM transactions (NATM).

Because the relationships among the partworth estimates obtained from the various methods and the observed behaviors are not necessarily linear, we used Spearman rank correlations to obtain the Multitrait Multi-method matrix. The resulting MTMM matrix is shown in Table 1, organized by methods. We also include the test-retest reliability correlations in the diagonals. These test-retest correlations show that the OLS and HAG methods tend to produce more reliable measurement than the other two approaches.
1.5 MTMM Analysis

Once creating the MTMM matrix in Table 1, we conducted a conventional MTMM analysis as suggested by Campbell and Fiske (1959) and a direct product model. A visual inspection of the MTMM matrix indicates some evidence of convergent and discriminant validity among the estimation methods. Convergent validity among the four estimations is clearly evident for MINBAL and CHECK, for which the monotrait-heteromethod correlations are very strong. Convergent validity for FEE is not as strong, but still statistically significant at 0.01. Convergent validity for ATMv is only evident for OLS, KAM and LCA.

Discriminant validity is also generally established in Table 1 among the four estimation methods, for MINBAL, CHECK and FEE. Monotrait-heteromethod correlations are larger than the heterotrait-heteromethod and heterotrait-monomethods correlations. However, discriminant validity (especially for the third criterion discussed earlier in this chapter, regarding the patterns of heterotrait correlations) is more easily verified with the direct product model, as discussed below.

In order to establish a stronger comparison of the different conjoint models, we applied a direct product decomposition to the MTMM matrix. We used MUTMUM (Browne 1992) to test the convergent and discriminant validity of the MTMM matrix. Table 2 presents the matrices of correlations among methods, PM, and traits, PT. As can be seen from the PM matrix, the correlations among the different conjoint methods are very large. More specifically, as can be seen from the matrix, the correlations are significantly different from zero and sufficiently large, indicating strong convergent validity among the estimation methods. However, when it comes to explaining actual behavior, the part-worth estimates of the conjoint models have relatively low correlations with the actual banking behavior. This implies that even though the conjoint models are consistent in their predictions, they have a relatively lower explanatory power with respect to actual behavior. This can be due to many external factors affecting actual behavior that were not included in the conjoint models. Among the four methods, OLS and HAG produce slightly better correlations with actual behavior, while LCA produces the lowest correlation, which is understandable, because the latter method constrains the individual-level estimates to the convex hull of the latent-class estimates.
Table 2. $PM$ and $PT$ matrices

a. $PM$: Method Correlations

<table>
<thead>
<tr>
<th></th>
<th>ACT</th>
<th>OLS</th>
<th>KAM</th>
<th>HAG</th>
<th>LCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.468</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAM</td>
<td>0.441</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAG</td>
<td>0.484</td>
<td>0.972</td>
<td>0.913</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>LCA</td>
<td>0.371</td>
<td>0.875</td>
<td>0.843</td>
<td>0.858</td>
<td>1.000</td>
</tr>
</tbody>
</table>

ACT: Actual behavior,
OLS: OLS estimation,
KAM: Kamakura's methodology,
HAG: Hagerthy's methodology, and
LCA: Latent Class Analysis.

b. $PT$: Trait Correlations

<table>
<thead>
<tr>
<th></th>
<th>MBAL</th>
<th>CHE</th>
<th>FEE</th>
<th>ATM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBAL</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHE</td>
<td>0.651</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEE</td>
<td>-0.579</td>
<td>-0.912</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>ATM</td>
<td>0.343</td>
<td>-0.111</td>
<td>0.168</td>
<td>1.000</td>
</tr>
</tbody>
</table>

MBAL: Minimum Balance,
CHE: Number of checks issued per month,
FEE: Number of times that the individuals paid a monthly fee, and
ATM: The number of ATM transactions.

On the other hand, a visual inspection of the $PT$ matrix indicates that the correlations among the different traits vary. For instance, consistent with our expectations, the amount of balance in the account is negatively correlated with the monthly fee paid (i.e., the less an individual has in his/her bank account, the more monthly fee he/she pays.) Similarly, the number of checks issued has a negative correlation with the number of monthly fee paid.

One advantage of the direct product model is that we can directly test the convergent and discriminant validity criteria established by Campbell and
Fiske (1959). Convergent validity is achieved when the correlations among the methods in the $P_M$ matrix are positive and large. From the $P_M$ matrix, it is evident that the correlations are positive and sufficiently large. Thus, the conjoint methods are in agreement with each other. Discriminant validity is achieved when (1) the correlations among traits in $P_T$ are low; (2) the method correlations in $P_M$ are greater than the trait correlations in $P_T$; and (3) the direct product model fits the data (Bagozzi and Yi 1991). As can be seen from the $P_T$ matrix, the first two conditions for discriminant are only partially met. The trait correlation between CHECK (with inverted sign) and $FEE$ is strong and negative, indicating that customers who are sensitive to charges per check are also sensitive to monthly fees. In order to assess the last condition for the discriminant validity, we looked at the fit indices for the direct product model. Based on the OLS estimation procedure of MUTMUM, the model’s discrepancy function value was 6.438 with corresponding $\chi^2 = 1884.31$, d.f. = 159 and $p = 0.00$. In addition, the model’s Root Mean Square Error of Approximation (RMSEA) was 0.201. All these indices suggest that the model does not fit the data. Overall, when we take into account all these three conditions we can conclude that the traits are not distinct. This is consistent with our initial expectations that the traits in this study should be correlated. Therefore, the lack of discriminant validity is a reflection of the traits being measured, rather than the measurement instruments.

Although we have theoretical and empirical reasons for using the direct product model for conducting the MTMM analysis, we also wanted to use the CFA to see whether we could get similar results. Like the models of many other researchers (Kalleberg and Kluegel 1975; Lee 1980; Marsh and Hocevar 1983; and Lastovicka et al. 1990), our model, after more than a dozen attempts, also yielded uninterpretable results, such as correlations outside the $-1$ and $+1$ range, negative unique variances, and nonconverging solutions. As suggested by Bagozzi and Yi (1991), these types of solutions can also be considered as an indication of why an additive model (CFA) is not appropriate to analyze MTMM matrices such as ours.
1.6 Conclusions and Directions of Future Research

Conjoint analysis has been a popular methodology for both researchers and practitioners. One important aspect of conjoint analysis is the validity of the results. As we discussed earlier, previous studies provide quite a large number of ways of improving the validity of conjoint analysis. Previous literature also presents a number of comparative studies investigating the extent to which those methods improve the validity of conjoint analysis. The results of these comparative studies have been inconclusive. The inconclusive results might be due to various reasons. One of the important reasons is the validity measure used for comparison. Our literature review indicates that previous studies have primarily used holdout samples and simulations. Only a few rare studies validated conjoint analysis with actual behavior, observed after a certain time period such as three months.

We used actual banking behavior of individuals immediately prior to a conjoint experiment as a benchmark for comparison, eliminating the concerns about the validity measures. By using an MTMM framework, we compared the actual banking behavior on four dimensions with the part worth estimates generated by the OLS procedure, Kamakura’s and Hagerthy’s methodologies, and the latent class analysis. The results were consistent with the findings of Vriens et al. (1996), indicating that the methods had more or less similar predictive performance. However, one interesting result was that despite the strong consistency among the methods, the correlations between the estimates and the actual behavior were relatively low, which is understandable due to the multitude of factors affecting actual behavior, but are not considered in the conjoint design.

Another aspect of the partworth estimates rarely considered in conjoint validation studies is their logical consistency. For example, in our study one should expect customers to prefer to keep a lower minimum balance, pay less per check, pay lower monthly fees, and have free access to ATM’s. Table 3 reports the percentage of the 269 respondents with logically consistent estimates for each of these attributes for the four methods we tested. One can see that all methods produced a high percentage of logically consistent estimates, but the segment-level models (KAM and LCA) performed slightly better.
Proportion of cases with logically consistent estimates

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>KAM</th>
<th>HAG</th>
<th>LCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Balance</td>
<td>92.6%</td>
<td>96.7%</td>
<td>90.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Cost/check</td>
<td>91.8%</td>
<td>95.2%</td>
<td>93.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Monthly Fee</td>
<td>94.4%</td>
<td>97.4%</td>
<td>98.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Free-Paid ATM</td>
<td>95.2%</td>
<td>92.9%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Overall</td>
<td>93.5%</td>
<td>95.5%</td>
<td>95.6%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Table 3 The percentage of the 269 respondents with logically consistent estimates for each of these attributes for the four methods tested.

This is certainly just one attempt to cross-validate preference measurements obtained through various conjoint models. Replication of the Multitrait Multi-method analysis for other applications of conjoint analysis, using real behavior as a criterion for external validity, are needed before generalizable conclusions can be drawn about the validity of preference measurements via conjoint analysis. In addition, future research can extend this study by investigating other conjoint models and other data collection modes.

1.7 References


© Springer


Leigh, Thomas W., David B. MacKay and John O. Summers (1984), "Reliability and Validity of Conjoint Analysis and Self-


