Reprinted with permission from the *Journal of Marketing Research*, published by the American Marketing Association, Oded Netzer, and V. Srinivasan, 48, no. 1 (Winter 2011): 140-156.

ODED NETZER and V. SRINIVASAN*

The authors propose a Web-based adaptive self-explicated approach for multiattribute preference measurement (conjoint analysis) with a large number (ten or more) of attributes. The proposed approach overcomes some of the limitations of previous self-explicated approaches. The authors develop a computer-based self-explicated approach that breaks down the attribute importance question into a ranking of attributes followed by a sequence of constant-sum paired comparison questions. In the proposed approach, the questions are chosen adaptively for each respondent to maximize the information elicited from each paired comparison question. Unlike the traditional self-explicated approach, the proposed approach provides standard errors for attribute importance. In two studies involving digital cameras and laptop computers described on 12 and 14 attributes, respectively, the authors find that the ability to correctly predict validation choices of the proposed adaptive approach is substantially and significantly greater than that of adaptive conjoint analysis, the fast polyhedral method, and the traditional self-explicated approach. In addition, the adaptive self-explicated approach yields a significantly higher predictive validity than a nonadaptive fractional factorial constant-sum paired comparison design.

Keywords: preference measurement, conjoint analysis, self-explicated methods, marketing research, multiattribute models

Adaptive Self-Explication of Multiattribute Preferences

Two important trends have significantly influenced the application of multiattribute preference measurement techniques (conjoint analysis) in recent years. First, as products and services become more complex and consumers become better informed about various features, marketers are measuring consumers' preferences over a larger number of attributes (Bradlow 2005; Netzer et al. 2008; Rao and Hauser 2004). Researchers have conducted applications of conjoint analysis for hotels, credit cards, and technological products using up to 50 product attributes (e.g., Krieger, Green, and Wind 2004; Wind et al. 1989). The second trend involves the move of the market research industry to Web-based data collection. For conjoint analysis applications, computerized data collection has enabled researchers to develop adaptive questionnaires—such as the adaptive conjoint analysis (hereinafter ACA; Johnson 1987) and the fast polyhedral method (hereinafter FPM; Toubia et al. 2003)—that maximize the information collected in each question. However, the advent of the Web has also produced a decrease in respondents' patience for long questionnaires. Consequently, even ACA methods are limited in their ability to handle complex problems with a large number of attributes and attribute levels. The self-explicated method (hereinafter SEM; Leigh, MacKay, and Summers 1984; Marder 1997; Srinivasan 1988; Srinivasan and Wyner 1989) is well suited to measure consumer preferences for multiattribute products that involve a large number of attributes, but it is often criticized for not capturing the trade-offs consumers face when making decisions (Green and Srinivasan 1990).

*Oded Netzer is the Philip H. Geier Associate Professor of Business, Columbia Business School, Columbia University (e-mail: on2110@columbia.edu). V. Srinivasan is Adams Distinguished Professor of Management, Graduate School of Business, Stanford University (e-mail: seenu@stanford.edu). The authors thank Pradeep Bansal, Yuval Efrati, Ping Li, Mingjie Lin, Anita Rao, Rom Schrift, Olivier Toubia, and seminar participants at Columbia University, Hebrew University, Marketing Modelers Group NYC, Marketing Science Institute, Stanford-Berkeley Marketing Colloquium, University of California–Los Angeles, University of Georgia, and Yale Center for Customer Insights for comments on previous versions of this research. John Hauser served as associate editor for this article.

Two methods are commonly used to estimate attribute importance in self-explicated studies: ratings and constant-sum allocation. A common problem with the ratings approach is that it does not explicitly capture the trade-offs between attributes; it is easy for respondents to say that every attribute is important. The constant-sum approach overcomes this limitation, but when the number of attributes is large (e.g., ten or more), it becomes difficult for a respondent to divide a constant sum among all the attributes. In this article, we propose an adaptive method, which we call adaptive self-explication (hereinafter ASE, pronounced "ace"), to overcome the limitations of traditional SEMs. Our approach breaks down the attribute importance question into a ranking of the attributes followed by a sequence of constant-sum paired comparison questions between two attributes at a time (not two partial product profiles at a time, as in ACA and FPM). The paired comparison questions are chosen adaptively for each respondent to maximize the information elicited from each question. The proposed approach provides improvement over current SEMs on several dimensions. First, unlike rating-scale methods, constant-sum paired comparison questions can capture trade-offs between product attributes. Second, by breaking down the attribute importance question into a ranking plus a subset of constant-sum paired comparison questions, researchers can study problems with a relatively large number of attributes, eliminating the difficulty of performing the constant-sum task across all attributes. Third, the adaptive nature of the questionnaire enables researchers to reduce the number of questions they ask each respondent and thus reduces respondents' burden and improves predictions. Finally, unlike the traditional SEM, the proposed approach provides standard errors for attribute importance.

We use two empirical applications to test the proposed approach. The first study involves preference for digital cameras described along 12 attributes to compare the predictive validity of the ASE approach with the ACA, FPM, and SEM. We estimate the methods using both traditional and hierarchical Bayes estimation. In the second application, we replicate the comparison of the ASE with the ACA for laptop computers varying along 14 attributes and extend the previous application by comparing the ASE with a non-adaptive fractional factorial constant-sum paired comparison design. This comparison enables us to assess the contribution of the adaptive design beyond the idea of breaking down the constant-sum importance question into a set of constant-sum paired comparison questions. Across both studies, we find a significant and substantial improvement in predictive validity for ASE over the alternative methods, even for a relatively short ASE questionnaire.

The next section describes our proposed approach. Then, we describe the two empirical applications and compare our proposed approaches with alternative preference measurement methods. We conclude with a discussion and directions for further research.

THE ASE APPROACH

Measuring Multiattribute Preferences for a Large Number of Attributes

Traditional preference measurement methods, such as the full-profile method (Green and Rao 1971) and choice-based conjoint analysis, are not suitable for problems involving a large number of attributes because respondents have difficulty processing a large number of attributes at a time (Green and Srinivasan 1990; Orme 2007). Researchers have suggested several approaches to address the increasing demand to measure preference along a large number of product attributes. Stated preference methods such as the SEM (Srinivasan 1988) are capable of handling a larger number of product attributes, but they carry their own limitations (Green and Srinivasan 1990). Other researchers have proposed hybrid methods that combine the self-explicated stage with graded paired comparisons of partial product profiles (e.g., hybrid conjoint analysis [Green, Goldberg, and Montemayor 1981; Marshall and Bradlow 2002; Ter Hofstede, Kim, and Wedel 2002], ACA [Johnson 1987]) and the partial profile choice-based conjoint method (Orme 2007) to overcome these problems. Evidence regarding their ability to improve predictive validity over full-profile analysis or a self-explicated task is mixed (e.g., Agarwal and Green 1991; Green, Krieger, and Agarwal 1991; Huber et al. 1993; Srinivasan 1988; Srinivasan and Park 1997). Scholz, Meissner, and Decker (2010) offer an alternative compositional approach to overcome the limitation of the traditional self-explicated approaches. Similar to the proposed approach, they use constant-sum paired comparison questions, but unlike our approach, they use the eigenvalue-based analytic hierarchy process (hereinafter AHP) to estimate the partworths and do not employ an adaptive design.

More recently, researchers have suggested machine learning approaches that employ complexity control to handle large-scale preference measurement tasks (Cui and Curry 2005; Evgeniou, Boussios, and Zacharia 2005; Hauser et al. 2010). Park, Ding, and Rao (2008) propose an innovative method for handling a large number of product attributes using an incentive-aligned product upgrading design. Table 1 summarizes the existing compositional and hybrid approaches to measure preference for products that have a large number of attributes.

In this article, we extend the traditional SEM. In the proposed approach, the self-explicated attribute importances are collected using an adaptively chosen sequence of constant-sum paired comparison questions. We first briefly describe the SEM and then detail our proposed adaptive approach to extend it.

The Traditional SEM

The SEM includes two data collection stages: ratings of the desirability of attribute levels within each attribute and ratings of the relative attribute importance across attributes (Srinivasan 1988; Srinivasan and Wyner 1989). In the first stage, respondents use a rating scale to evaluate how desirable each of the levels of each product attribute is to them. In the second stage, the survey prompts respondents to evaluate how important each attribute is to them. Specifically, respondents are asked to evaluate how valuable the improvement from the least to the most preferred level of each attribute is to them (Srinivasan 1988). Sawtooth's ACA software currently implements this question format. Regardless of the method used to obtain the within- and across-attributes self-explicated data, the ACA method calculates the self-explicated partworths as follows:

1. Rescale the desirability ratings such that the most preferred attribute level of each attribute for that respondent receives a

Table 1

SELF-EXPLICATED AND HYBRID PREFERENCE MEASUREMENT APPROACHES

Method	*Description*	*Comments*
SEMs (Leigh, MacKay, and Summers 1984; Marder 1997; Srinivasan 1988; Srinivasan and Wyner 1989)	Researchers ask the respondent directly about the importance he or she places on each attribute and his or her desirabilities for all attribute levels within each attribute.	•A low cognitive load task, which provides surprisingly robust predictions (Srinivasan and Park 1997). •The method is susceptible to biases due to its self-reporting nature (see Green and Srinivasan 1990 for a review).
Hybrid conjoint analysis (Green, Goldberg, and Montemayor 1981; Marshall and Bradlow 2002; Ter Hofstede, Kim, and Wedel 2002)	This approach combines the self-explicated approach with overall evaluation of a small number of full profiles.	•Evidence regarding improvement over the SEM is mixed.
ACA (Sawtooth Software, Johnson 1987)	This method combines the self-explicated approach with graded paired comparisons of partial profiles. Paired comparison questions are chosen adaptively to maximize utility balance.	•One of the most commonly used approaches to measure preferences for a large number of attributes. •Evidence regarding improvement over the SEM is mixed.
FPM (Toubia et al. 2003)	This method combines the self-explicated approach with graded paired comparisons of partial profiles chosen adaptively on the basis of the analytic center approach.	•Improved predictive ability relative to ACA.
Paired comparison-based preference measurement (Scholz, Meissner, and Decker 2010)	The researchers ask a series of constant-sum paired comparison questions estimated using the AHP.	•Improved predictive ability relative to the SEM and ACA. •Unlike the ASE, this approach does not provide standard errors for the parameter estimates.
Web-based upgrading method (Park, Ding, and Rao 2008)	Respondents are endowed with a product profile and then are allowed to upgrade one attribute at a time using an incentive-aligned procedure.	•The method is incentive aligned. •Improved predictive ability relative to traditional SEM. •Each respondent provides information only about a subset of the attributes/attribute levels. Researchers use hierarchical Bayes to pool information across respondents.
ASE (This study)	Replaces the importance measurement stage of the SEM by ranking attribute improvements and adaptive constant-sum paired comparisons of attribute improvements.	•Improved predictive ability relative to traditional SEM, ACA, and FPM. •The method adds trade-offs between attributes to the traditional SEM.

Notes: Only the SEM is easily scalable to a large number of attributes. The other methods are moderately scalable to a large number of attributes.

rating of 10 and the least preferred level of each attribute for that respondent receives a rating of 0. Rescale the intermediate levels of each attribute accordingly. Let D_{jk} be the rescaled desirability ratings (on a scale of 0–10) for level k ($k = 1, 2, \ldots, K_j$) of attribute j ($j = 1, 2, \ldots, J$).[1]

2. Rescale (if necessary) the attribute importance ratings such that for each respondent, they sum to 100 across attributes. Let W_j denote the rescaled importance of attribute j such that $\Sigma_{j=1}^{J} W_j = 100$ for each respondent.
3. To obtain the self-explicated partworths, multiply the importance measures by the rescaled desirability ratings and divide by the range of the desirability scale:

$$P_{jk} = W_j \times D_{jk}/10. \tag{1}$$

The most commonly used scale to measure attribute importance is a rating scale, in which respondents rate how valuable the improvement from the least to the most preferred level of the attribute is to them. There are several reasons for the popularity of the rating scale approach: First, ratings are relatively easy for respondents to complete (Krosnick 1999); second, they allow for incorporating a large number of attributes; and finally, they are easy to analyze. However, there are limitations with collecting attribute importance using a rating scale. First, because there is no trade-off involved, respondents tend to state that all attributes are important (Krosnick and Alwin 1988), leading to a relatively narrow distribution of attribute importance. Second, attribute importance is a ratio-scaled construct (the zero point refers to an irrelevant attribute). As such, ratio scale is more appropriate for collecting importance data.

One possible ratio-scaled approach to collecting attribute importances is the constant-sum scale, on which respondents are asked to allocate, for example, 100 points across the different attributes to reflect the relative importance of each attribute to them. This approach minimizes the problems we mentioned previously but introduces a new problem: Even the most diligent respondent finds it difficult to allocate a constant sum across a large number of attributes (e.g., ten or more). The proposed approach helps alleviate this problem by breaking down the constant-sum question across all attributes into a rank order of the attributes followed by a sequence of constant-sum paired comparison questions. The essence of the proposed approach is to improve the estimation of the individual-level attribute importances (W_j), while maintaining the general framework of the self-explicated approach, in a way that introduces trade-offs but avoids respondent overload.

[1]In all the methods studied here, we collect the data and estimate all parameters at the individual level. For ease of exposition, we left out the subscript i for individuals ($i = 1, \ldots, I$).

Adaptive Constant-Sum Paired Comparisons Design

The proposed approach breaks down the cognitively demanding constant-sum allocation across the full set of attributes into a subset of constant-sum allocations between two attributes at a time. Previous research has found that the constant-sum paired comparison approach is a reliable method to capture preferences (Axelrod 1968; Saaty 1980), and Silk and Urban (1978) use it successfully in the Assessor model. A possible difficulty with breaking down the constant-sum allocation question into a set of paired comparison questions is that when the number of product attributes becomes large, the number of possible paired comparison questions increases significantly. More specifically, there are $J \times (J - 1)/2$ possible pairs, where J is the number of product attributes. Thus, even for a problem with ten product attributes, there are 45 possible paired comparison questions. To reduce the number of paired comparison questions, we propose ranking the attributes together with an adaptive approach. The general idea behind our adaptive design is to ask the respondent to compare only a subset of all possible paired comparison questions, chosen adaptively to maximize the information each question elicits.

Similar to the first stage in the traditional self-explicated approach, the first stage of the proposed approach elicits a set of desirability ratings, one for each attribute level (for a screen shot of a desirability ratings scale, see Figure 1). In the second stage, the respondent ranks all the product attributes according to their importance to him or her (for a screen shot of the attribute importance ranking task, see Figure 2). The third stage includes a set of constant-sum paired comparison questions. For each question, the respondent allocates 100 points, using a sliding bar, initially located at equal importance to each of the two attributes, to reflect the relative importance placed on the improvement from the least to the most preferred levels of two attributes (for a screen shot of a constant-sum paired comparison question, see Figure 3).[2] Both ASE and ACA use paired comparisons. However, it is important to note that in ASE the paired comparisons are comparisons of attributes, whereas in ACA (and FPM), they are paired comparisons of partial product profiles. If a researcher wants to estimate the likelihood of purchase in the product category, he or she can add a fourth stage that includes a set of purchase intention questions for several (e.g., five) product profiles (similar to the approach used in ACA). The researcher can then estimate the attribute importances and the corresponding (approximate) standard errors from the constant-sum paired comparison questions using a log-linear regression. Next, we describe the log-linear regression and then describe the adaptive algorithm used for data collection.

Estimating the attribute importances and partworths. The output from the constant-sum paired comparison questions are ratios of the importances of two attributes at a time. Given the set of attribute importances ratios, we use a log-linear multiple regression to estimate attribute importances. We outline the proposed log-linear estimation procedure as follows:

[2]If there was a tie in the desirability ratings for the lowest- or highest-rated levels, we randomly chose one of the tied levels.

Figure 1
THE DESIRABILITY RATINGS USER INTERFACE

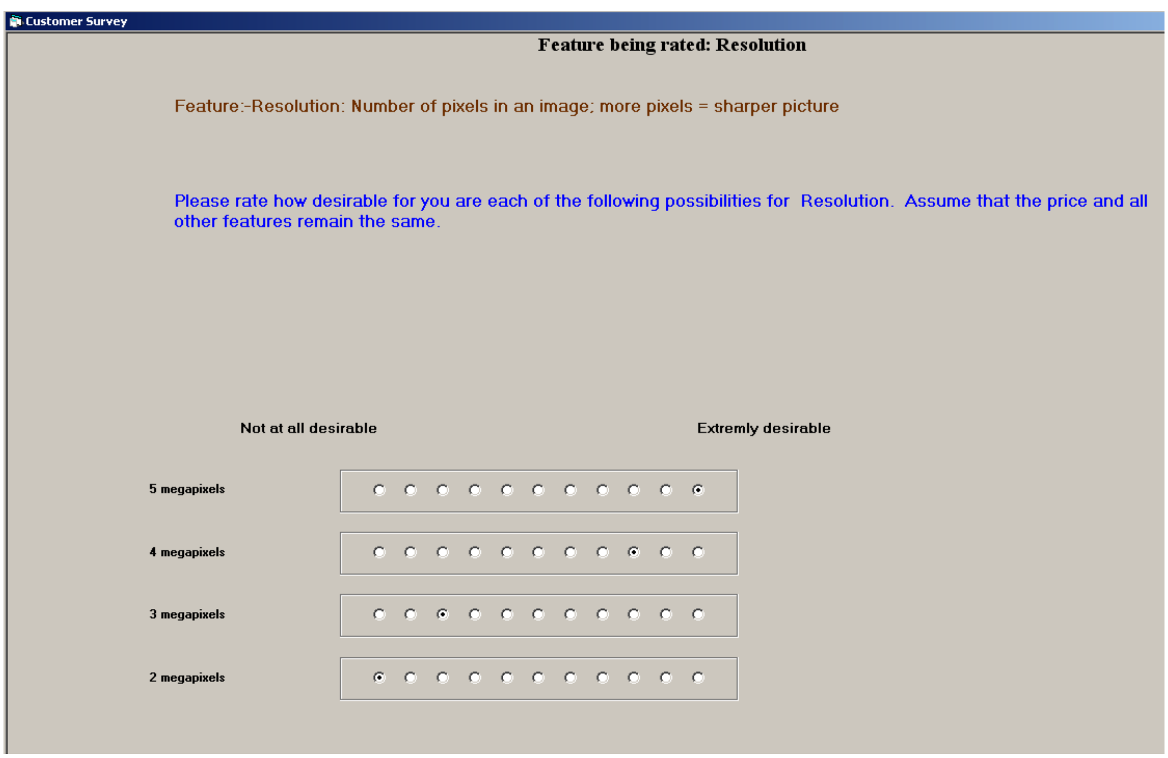
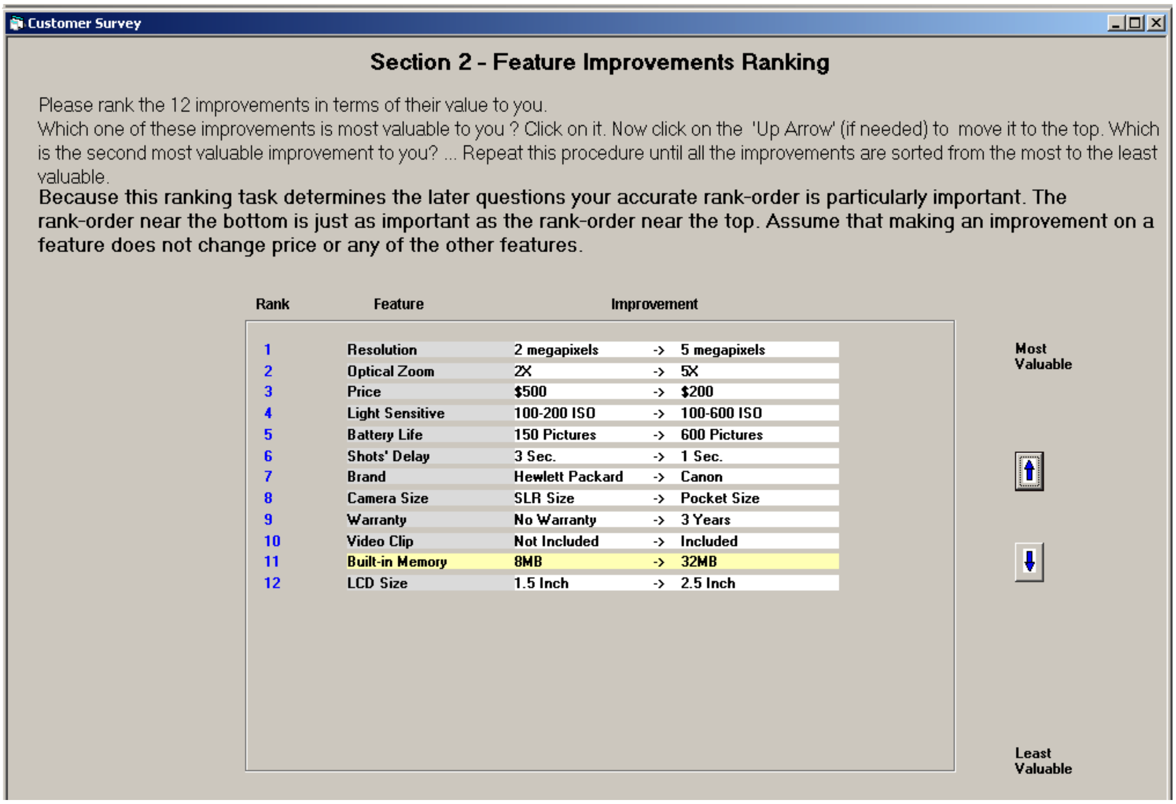


Figure 2
THE ATTRIBUTE IMPORTANCE RANKING TASK USER INTERFACE



1. Let W_j be the importance of attribute j.
2. The ratio of the importance of attribute j_1 to attribute j_2 is $r_{j_1j_2} = W_{j_1}/W_{j_2}$.[3]
3. Let $V_j = \log(W_j)$, where the log is taken to the base 10 (without loss of generality). Thus, $V_{j_1} - V_{j_2} = \log(r_{j_1j_2})$.
4. Without loss of generality, relabel the attributes so that 1 denotes the most important attribute in the ranking for this respondent. Let j = 1, 2, …, J denote the attributes involved in the constant-sum paired comparison questions.
5. Without loss of generality, set $V_1 = V_{highest} = a$, where a is a positive number (e.g., a = 2 corresponds to $W_1 = 100$).
6. Define Ω as an N × (J – 1) design matrix, where the columns correspond to attributes j = 2, 3, …, J, and N is the number of paired comparison questions. For identification purposes, we drop the first (most important) attribute from the design matrix. Note that each paired comparison has top and bottom attributes (see Figure 3). Each row n of the design matrix corresponds to a paired comparison question such that for j = 2, 3, …, J,

$$\Omega(n, j) = \begin{Bmatrix} 1 \text{ if } j \text{ is the top attribute in the pair } n \\ -1 \text{ if } j \text{ is the bottom attribute in the pair } n \\ 0 \text{ if attribute } j \text{ is not in the pair } n \end{Bmatrix}.$$

7. Define **R** as a vector of N × 1 log ratios such that each element in the vector is log ratio of the corresponding paired comparison question: $\log(r_{j_1j_2})$. Following Step 5, if $j_1 = 1$, replace $\log(r_{1j_2})$ with $\log(r_{1j_2}) - a$; if $j_2 = 1$, replace $\log(r_{j_11})$ with $\log(r_{j_11}) + a$.
8. Run a log-linear (multiple) regression of **R** on the design matrix Ω. From the multiple regression, obtain a set of individual log importances, V_j, for j = {2, 3, …, J}.
9. Take antilog of V_j, $W_j = 10^{V_j}$. Recall that $W_1 = 10^a$.
10. Normalize the importances $\{W_j\}$ such that $\Sigma_{j=1}^{J} W_j = 100$.
11. Plug in the rescaled W_j and the desirability ratings D_{jk} into Equation 1 to obtain the individual-level partworth functions.

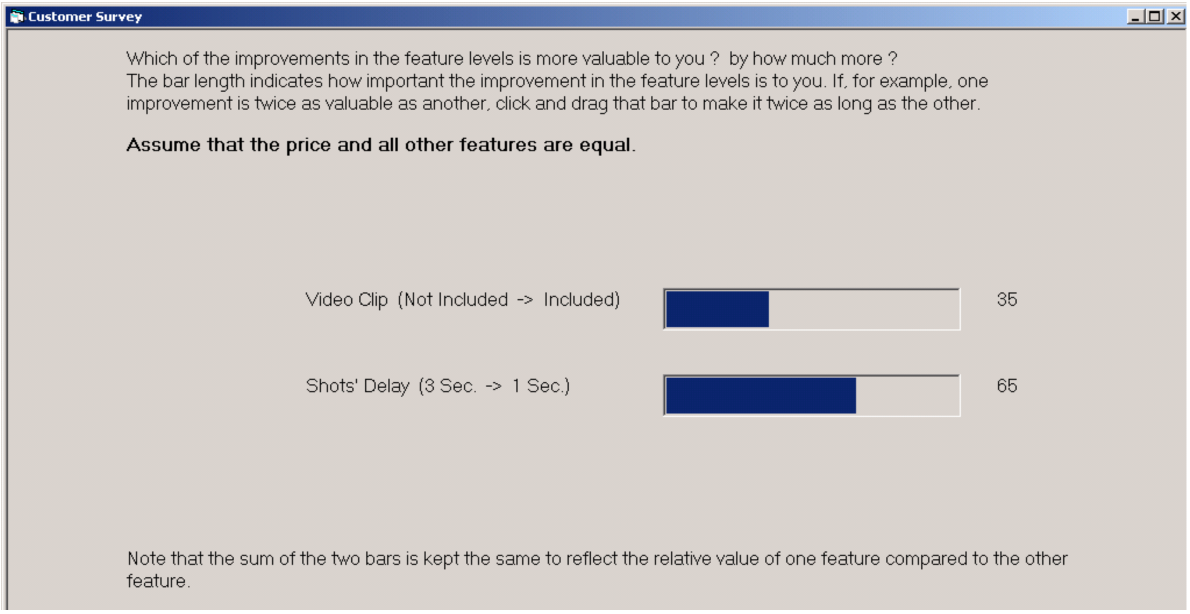
Unlike the traditional SEM approach, we derive the attribute importances in our approach from an ordinary least squares regression, and therefore we estimate them with standard errors. To translate the ordinary least squares standard error of $V_j(\hat{\sigma}_{V_j})$ to that of $W_j(\hat{\sigma}_{W_j})$, we use the Taylor series approximation:

$$(2) \qquad \hat{\sigma}_{W_j} = \hat{\sigma}_{10^{V_j}} \approx \ln(10)10^{V_j}\hat{\sigma}_{V_j}.$$

An alternative approach to estimate importances from constant-sum paired comparison questions data is the AHP (Saaty 1980). The AHP calculates importances using an eigenvector approach. A limitation of Saaty's (1980) AHP approach, relative to the proposed log-linear regression, is that the eigenvector AHP approach requires asking all possible paired comparison questions. Furthermore, unlike our regression approach, the AHP approach does not provide standard errors for the attribute importance estimates; it only provides an overall measure of inconsistency of the pairs. Researchers have proposed a few approaches to mod-

[3]In the rare situation that the respondent's allocation is 100:0, we avoid division by zero by converting it to 97.5:2.5.

Figure 3
THE CONSTANT-SUM PAIRED COMPARISON QUESTION USER INTERFACE



ify the AHP to allow for missing pairs, thus reducing the number of needed paired comparisons (Harker 1987; Scholz, Meissner, and Decker 2010; Weiss and Rao 1987). However, these approaches either produce only aggregate-level estimates across decision makers or are limited in their ability to reduce the number of paired comparison questions without significant loss in accuracy (Carmone, Kara, and Zanakis 1997). For a relatively small-scale problem of an MBA student job choice study involving eight attributes, we compared the attribute importances estimated from our proposed log-linear regression and the AHP eigenvector approach and obtained an average correlation of .98 (SD = .034) and cosine theta of .99 (SD = .020) between the two approaches, demonstrating high convergent validity between the two methods.

The 11 steps listed previously detail how to estimate attribute importances and partworths from the set of constant-sum paired comparison questions using a log-linear regression. The key remaining question is how to choose the subset of constant-sum paired comparison questions to be asked. We describe our adaptive approach for paired comparison question selection next.

Adaptive data collection. The proposed approach asks only a subset of all possible constant-sum paired comparison questions. We use the ranking information from the second stage of data collection to interpolate the importance of the attributes not included in the subset of paired comparison questions. This interpolation creates error, and the adaptive methodology selects the attributes for the next paired comparison question to minimize the maximum sum of interpolation errors at that stage (which we elaborate on next). Another consideration is that we need to assess the consistency of the respondent in providing the constant-sum paired comparisons. We assume that the constant-sum response data are ratio scaled. To assess the validity of our assumption of ratio scaled responses, we built in some redundancy in the paired comparison questions. The log-linear regression's adjusted R-square provides a measure of ratio-scaled consistency of the respondent's data. The redundant paired comparisons also permit us to determine (approximate) standard errors for the estimated importance ratings. The redundancy in paired comparison questions puts us well within the 1.5:2 ratio of observations (paired comparisons) to estimated parameters (weights) commonly used in conjoint analysis studies.

The adaptive design consists of the following steps:

1. The survey prompts the respondent to rank all the attributes according to their importance (i.e., value of the improvement from the least to the most preferred level for that respondent; see Figure 2). The initial order of attributes on the ranking screen is chosen randomly for each respondent to minimize order bias at the aggregate level. If the number of product attributes is large (e.g., ten or more), first, the survey prompts the respondent to split the set of attributes into two (or more) categories of more and less important attributes. Second, the respondent ranks attributes within each of the categories. At the end of the ranking task, the attributes are relabeled so that 1 denotes the most important attribute in the ranking task, 2 denotes the second most important attribute, and so on; J denotes the least important attribute.
2. Using the ranking data, the proposed method prompts the respondent to answer three constant-sum paired comparison questions (see Figure 3). The three questions compare the attribute ranked first with the attribute ranked last, the attribute ranked first with the attribute ranked middle—that is, the attribute ranked (J + 1)/2 if J is odd (or the attribute ranked J/2 if J is even)—and the attribute ranked middle with the attribute ranked last.
3. Using the log-linear multiple regression we described previously, estimate the attribute importance of the attributes ranked first, middle, and last. That is, estimate W_1, $W_{(J + 1)/2}$, W_J.

4. At this stage, there are two intervals (i.e., subsets of attributes): the subset intermediate between the first- and middle-ranked attributes, denoted as {1, (J + 1)/2}, and the other intermediate between the middle- and last-ranked, denoted as {(J + 1)/2, J}. (We have not yet estimated the importance of attributes in these intervals.) Each interval has a top and a bottom attribute. For example, in the interval {1, (J + 1)/2}, the top attribute is 1 and the bottom attribute is (J + 1)/2. In the absence of any additional information, our best estimate for the importance of the intermediate attributes would be a linear interpolation between the importance of attribute 1 and attribute (J + 1)/2 based on the rank order. This interpolation creates error. (It should be noted that this error is due to interpolation of attributes not asked and is different from the estimation error of importance estimates in the log-normal regression.) In choosing which interval to explore, we use the criterion of minimizing the maximum possible sum of interpolation errors. That is, we "open" the interval for which the worst possible sum of interpolation errors from using linear interpolation of the intermediate attributes would be the largest. In the example Figure 4 depicts, in Scenario A, we would open Interval I, and in Scenario B, we would open Interval II. More formally, we can visualize the maximum possible sum of interpolation errors as the area of the triangle between the linear interpolation and the horizontal and vertical lines defined by the top and bottom attributes (see Figure 5). Although this result may seem to be an approximation based on geometry, it is an exact result and can be analytically shown by summing the interpolation errors shown by the vertical lines in Figure 5. The maximum possible sum of errors is given as follows:

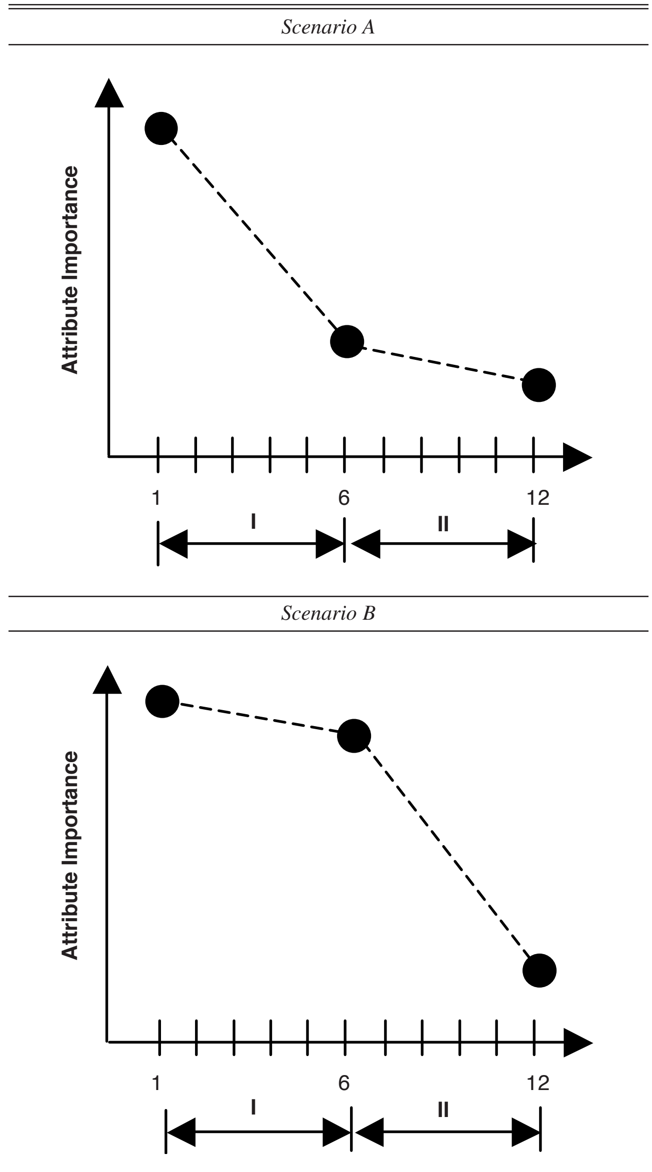
$$\text{Maximum possible sum of errors} = (W_{top} - W_{bottom}) \times (\text{number of intermediate attributes})/2.$$

Thus, the next interval to be opened would be that with the maximal $(W_{top} - W_{bottom}) \times$ (number of intermediate attributes). We recognize that this procedure minimizes the maximum sum of interpolation errors at each iteration and does not necessarily produce global optimality over the whole procedure; however, with current computing resources, using dynamic programming would be computationally too time consuming for Web-based real-time data collection.
5. When an interval is selected to be opened, the middle attribute in the interval is chosen as the attribute to measure importance (Again, the logic is one of minimizing the maximum possible sum of interpolation errors.) Two additional paired comparison questions are asked, one comparing the attribute at the top of the interval with the attribute at the middle of the interval and one comparing the attribute at the middle with the attribute at bottom of the interval. These two paired comparisons are added to all the previously collected paired comparisons for this respondent, and the log-linear regression is reestimated to obtain the importance of the previous attributes plus the newly chosen attribute. Although one of these two paired comparisons is sufficient for the purpose of estimating the importance of the new attribute, we ask both comparisons so that the ratio-scaled consistency of the respondent's data can be determined by the adjusted R-square of the log-linear regression. The redundant data also help us estimate the (approximate) standard errors of the estimated importances, as detailed previously.
6. Steps 3–5 are repeated iteratively until the number of preset paired comparison questions have been reached or until

$$\frac{W_{top} - W_{bottom}}{\hat{\sigma}_{W_{top} - W_{bottom}}} \leq \alpha, \tag{3}$$

Figure 4

AN ILLUSTRATIVE EXAMPLE OF SELECTING AN INTERVAL TO OPEN



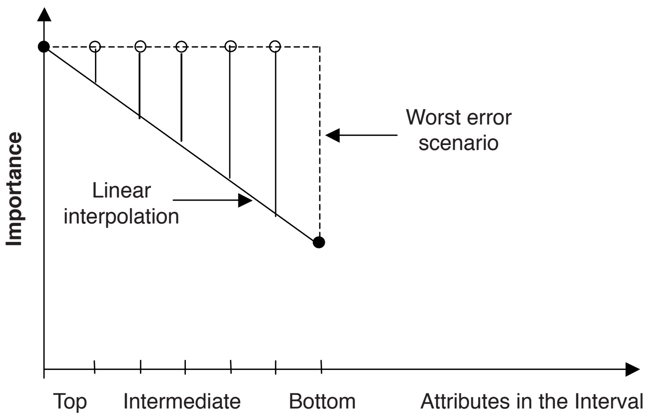
where $\hat{\sigma}_{W_{top} - W_{bottom}}$ is the standard error of $W_{top} - W_{bottom}$ calculated using the Taylor expansion:

$$\begin{aligned} \hat{\sigma}_{W_{top} - W_{bottom}} &= \hat{\sigma}_{10^{V_{top}} - 10^{V_{bottom}}} \\ &\approx \text{sqrt}\Big\{\Big[\ln(10)10^{V_{top}}\hat{\sigma}_{V_{top}}\Big]^2 \\ &+ \Big[\ln(10)10^{V_{bottom}}\hat{\sigma}_{V_{bottom}}\Big]^2 \\ &- 2[\ln(10)]^2 10^{V_{top}} 10^{V_{bottom}} \hat{\sigma}_{V_{top}, V_{bottom}}\Big\}, \end{aligned} \tag{4}$$

where $\hat{\sigma}_{V_{top}}$, $\hat{\sigma}_{V_{bottom}}$, and $\hat{\sigma}_{V_{top}, V_{bottom}}$ are the standard errors and covariance from the output of the log-linear regression and α is a user-defined factor. Note that this stopping criterion follows the same structure and rationale of the familiar t-test.
7. For attributes not included in the paired comparisons, importances are linearly interpolated according to the ranks using

Figure 5

THE MAXIMAL POSSIBLE SUM OF ERRORS CRITERION



the attributes at the top and bottom of the corresponding interval.[4] The (approximate) standard errors for the interpolated attributes are computed according to the standard errors (and covariance) of the corresponding top and bottom attributes.

Figure 6 depicts an example of the iterative paired comparison questions selection for one respondent. In the first stage, three paired comparison questions were asked, comparing the highest-ranked attribute (resolution) for this respondent with the lowest-ranked attribute (video clip), the highest-ranked with the middle-ranked (battery life), and the middle-ranked with the lowest-ranked. Using log-linear regression, we estimate the importance for resolution, battery life, and video clip. Figure 6 reports these numbers in the first column of numbers (scaled in such a way that the importance for the three attributes together with those of the remaining nine attributes obtained through interpolation sums to 100). Then, the adaptive algorithm evaluated which interval to open next (i.e., resolution–battery life or battery life–video clip). The first interval has a $W_{resolution} - W_{battery\_life} = 21.36 - 8.48 = 12.88$, and the number of intermediate attributes is four. For the second interval, $W_{battery\_life} - W_{video\_clip} = 8.48 - .70 = 7.78$, and the number of intermediate attributes is five. Thus, following the interval choice criterion in Step 4, the algorithm opened the interval resolution–battery life. The algorithm chooses price (the attribute at the middle of that interval) as the next attribute, resulting in two paired comparison questions: [resolution–price] and [price–battery life]. Then, the algorithm uses log-linear regression to estimate simultaneously the importance ratings of all four attributes from the five paired comparisons collected thus far. Figure 6 reports these numbers in the second column. At the next iteration, there are three intervals: (1) resolution–price, (2) price–battery life, and (3) battery life–video clip. Although Interval 1 has the largest gap between the top and the bottom in terms of importances, Interval 3 has the largest number of intermediate attributes. Following the interval choice criterion in Step 4, the algorithm chose Interval 3. This procedure is repeated until the stopping criterion in Step 6 is reached.

In adaptive questionnaires, it is crucial that the parameters estimates can be estimated in real time to avoid delay in the questionnaire progress. Because the ASE estimation procedure involves a simple log-linear multiple regression, the estimation of attribute importances at each iteration is extremely fast, and the respondent is not aware of the computation being done in the background. Next, we describe two empirical applications used to test the predictive ability of the proposed ASE approach.

EMPIRICAL APPLICATION 1: DIGITAL CAMERAS WITH 12 ATTRIBUTES

In the first empirical application, we use the context of digital cameras to compare the ASE with several commonly used preference measurement methods suitable for a large number of attributes. The digital camera category is important to the respondents' population (mostly students). Furthermore, consumers often consider a large number of attributes when purchasing a digital camera. Online retailers such as Yahoo Shopping and BestBuy.com describe digital cameras along more than 40 attributes. To keep our empirical application meaningful and realistic, we conducted a pretest to choose a set of 12 attributes that respondents found most important (for a list of the 12 attributes and their corresponding levels, see Table 2).

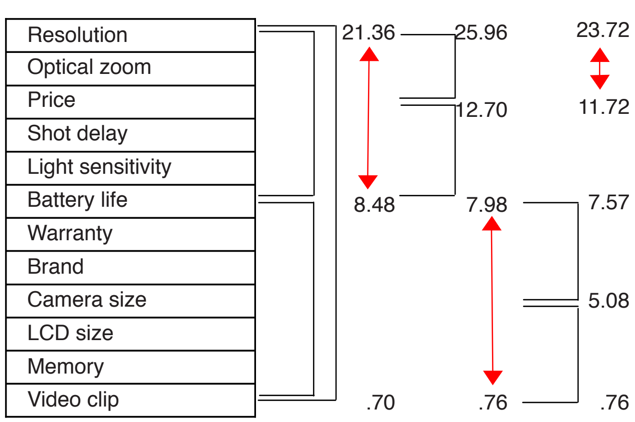
Research Design

We compared the ASE with the commonly used commercial ACA as well as the recently developed adaptive FPM (Toubia, Hauser, and Garcia 2007; Toubia, Hauser, and Simester 2004; Toubia et al. 2003). We recruited 154 participants through the behavioral lab of a West Coast university. We assigned them randomly to one of three preference measurement conditions: ASE (n = 52), ACA (n = 49), and FPM (n = 50).[5] The three groups were not statistically sig-

[4]In the unlikely event that the estimated importance of the attribute at the top of the interval is smaller than that of the attribute at the bottom of the interval, we set the importance of all the intermediate attributes in that interval to be the average of the top and bottom importance.

Figure 6

AN EXAMPLE OF THE ADAPTIVE PROCESS FOR SELECTING THE FIRST SEVEN PAIRED COMPARISON QUESTIONS



Notes: The lines connect the paired comparison questions asked at each stage. The arrows represent the interval to be opened in the next step. The numbers are the estimated importance ratings following all previous paired comparison questions. The numbers are scaled in such a way that together with interpolated values for the remaining attributes, they add to 100.

[5]We omitted three respondents because of incomplete survey data.

Table 2

ATTRIBUTES AND LEVELS FOR THE DIGITAL CAMERAS STUDY (EMPIRICAL APPLICATION 1)

Attribute	*Levels*
1. Brand	Canon, Hewlett-Packard, Nikon, Olympus, Sony
2. Battery life	150, 300, 450, 600 pictures
3. Built-in memory	8 MB, 16 MB, 32 MB
4. Camera size	Pocket size, medium size, SLR size
5. LCD size	1.5, 2, 2.5 inches
6. Light sensitivity	100–200, 100–400, 100–600 ISO
7. Optical zoom	2×, 3×, 4×, 5×
8. Price	\$500, \$400, \$300, \$200
9. Resolution	2, 3, 4, 5 megapixels
10. Shot lag	3, 2, 1 seconds
11. Video clip	Not included, included
12. Warranty	No warranty or 1-, 2-, or 3-year warranty

nificantly different in terms of age, gender, ownership of digital cameras, and familiarity with digital cameras ($p >$ .1). In all the conditions, respondents first completed a validation task, which we describe subsequently. After the validation task, respondents completed the preference measurement task (for the research design in the three conditions, see Table 3). We conducted the validation task before the preference measurement task to avoid any possible effect of the preference measurement task on the validation task. The reverse bias of the validation task influencing the preference measurement is minimal because all respondents completed the relatively time-consuming desirability ratings task before completing the preference measurement task. (In the second empirical application, we conducted both pre- and postvalidation tasks.) Finally, all respondents completed a postsurvey evaluation task.

In the ASE task, respondents first completed 12 desirability rating tasks, one screen for all the levels of a single attribute (see Figure 1) followed by the ranking task (see Figure 2). Next, the adaptive algorithm asked each respondent up to 21 constant-sum paired comparison questions (see Figure 3). We concluded the adaptive procedure when the ratio on the left-hand side of Equation 3 was zero or negative for all unopened intervals (i.e., $\alpha = 0$) or when there were no intervals to open. The average number of paired comparison questions asked per respondent was 17.8 with a maximum of 21 and a minimum of 13.

Respondents in the ACA and FPM conditions first completed 12 desirability rating tasks, similar to the ones used in the ASE, followed by an attribute importance rating task in which respondents rated, on a seven-point scale, the importance to them of the improvement from the least to the most preferred level of each attribute. These two steps constitute the traditional rating-based SEM questionnaire. We used these data separately to compare the ASE with the SEM. After the SEM task, respondents in the ACA and FPM conditions rated their relative preferences, on a nine-point scale, between 21 pairs of partial product profiles (for screen shots of the ACA and FPM interface, see the Web Appendix at http://www.marketingpower.com/jmrfeb11). Per Sawtooth Software's recommendation, the first 10 paired comparison questions included profiles varying along two attributes and the last 11 questions included profiles varying along three attributes. We chose 21 pairs for ACA and FPM to be consistent with ASE. This number is also within the range recommended for ACA surveys. To give the FPM paired comparison questions more power, we did not allow paired comparison questions in which one alternative strictly dominated the other. Similar to the common practice in ACA surveys, following the paired comparisons stage, we included in all three methods a set of five purchase intention questions to convert the partworths' overall utilities to purchase likelihood.

For the ACA questionnaire design and estimation, we used Sawtooth's SSI-Web Version 3.5.0. We used the open source code (http://mitsloan.mit.edu/vc) to implement the FPM adaptive survey and estimate the partworths using the analytical center approach.[6] The ACA and FPM provided similar respondent interfaces; the only difference between the two methods is the adaptive mechanism used to choose the paired comparison questions. It should be noted that in all methods, we used the partworth model (Green and Srinivasan 1978) and did not impose any restrictions on the partworth functions.

The validation task. The validation task included ranking of the alternatives in each of two choice sets of four digital cameras, each described on all 12 attributes. The procedure we employed to choose the validation choice sets ensured that each choice set would be Pareto optimal (i.e., none of the cameras within a choice set dominated any other). We randomly sampled 100 choice sets of four alternatives each

[6]We thank Professor Olivier Toubia of Columbia University for his help with adapting the FPM code to our empirical application.

Table 3

EXPERIMENTAL DESIGN (EMPIRICAL APPLICATION 1)

ASE	Two validation choice sets	Attribute-level desirabilities	Rank order of attribute importances	Paired comparisons of attribute importances	Purchase intention for five product profiles	Postsurvey evaluation
ACA/FPM	Two validation choice sets	Attribute-level desirabilities	Attribute importance ratings	Paired comparisons of partial product profiles on two and three attributes	Purchase intention for five product profiles	Postsurvey evaluation
SEM	Same validation data as in ACA/FPM	Partworths calculated on the basis of desirability and importance ratings of the ACA and FPM. No new data collected for SEM.		N.A.	N.A.	N.A.

Notes: N.A. = not applicable.

from a 64-profile fractional factorial design. Then, we eliminated choice sets in which there were more than four ties of attribute levels for at least one pair of alternatives and choice sets that had at least one identical attribute level across the four alternatives. From the remaining choice sets, we chose the four choice sets with the minimal Kendall tau distance. We calculated the Kendall tau distance for each pair of alternatives A and B by $\tau = |\Sigma^J_{j=1} I(L_{jA} > L_{jB}) - I(L_{jA} < L_{jB})|/[J - \Sigma^J_{j=1} I(L_{jA} = L_{jB})]$, where J is the number of ordinal attributes, L_{jA} and L_{jB} are the attribute levels of attribute j for alternatives A and B, respectively, and I is an indicator function. We calculate the choice set's overall Kendall tau distance by averaging the Kendall tau distance across the six pairs of alternatives in each choice set of four options. The average Kendall tau distance for the four chosen validation choice sets was .122. The survey prompted each respondent to rank the four alternatives in terms of their preferences in each of two choice sets randomly chosen from the four possible choice sets. We used the same validation task across the three preference measurement method conditions.

Predictive Validity

Individual-level predictions. Table 4 reports the individual-level predictive validity measures for the ASE, ACA, FPM, and SEM. Because we did not find significant order effect for the two validation choice sets, we aggregated the results across the two validation choice sets for each respondent. We used three measures of individual-level predictive validity. First, we measured the hit rates of predicting the highest-ranked alternative in each of the two validation sets (denoted by "choice set hit rate"). This measure could be thought of as the hit rates of the "chosen" alternative and thus is the most relevant validity measure for predicting choice. The second measure is the hit rate for the 12 pairwise choices derived from the ranking of four alternatives in the two choice sets (denoted by "pairs hit rates"). To account for the dependency between the two choice sets each respondent evaluated and the six pairwise choices derived from the ranking of each choice set of four options, we first average for each respondent the hit rates across the 12 pairs and then test the difference between the average hit rates across respondents in the different methods. Thus, the unit of analysis for the statistical test is a respondent. The third measure is the average rank-order correlation between predicted and actual ranking across the two validation sets for each respondent.

Table 4

COMPARISON OF INDIVIDUAL-LEVEL PREDICTIVE VALIDITY (EMPIRICAL APPLICATION 1)

	Choice Set Hit Rates		*Pairs Hit Rates*		*Rank-Order Correlation*	
Method (Sample Size)	*M*	*SE*	*M*	*SE*	*M*	*SE*
ASE (n = 52)	.606	.050	.718	.024	.517	.053
ACA (n = 49)	.398*	.044	.641*	.024	.351*	.057
FPM (n = 50)	.440*	.049	.655*	.020	.384*	.048
SEM (n = 99)	.449*	.034	.666*	.015	.397*	.036
ASE/HB (n = 52)	.644	.052	.727	.024	.515	.056
ACA/HB (n = 49)	.408*	.050	.658*	.024	.369*	.053
FPM/HB (n = 50)	.480*	.051	.687*	.020	.454*	.047

*Statistically significantly lower than the corresponding ASE or ASE/HB results ($p < .05$).

The ASE was able to correctly predict the highest-ranked alternatives in 61% of the choice sets and the pairs hit rates in 72% of the pairs. The ASE hit rates and rank-order correlations between the predicted and actual ranking of the validation choice sets were substantially and significantly higher ($p < .05$) than those of the ACA and the FPM. The ASE provided 35%–52% improvement in validation choice set hit rates over the alternative methods. The hit rates and rank-order correlations of the ASE were also substantially and significantly higher than those of the SEM ($p < .05$), suggesting that ranking the attributes, breaking down the importance question into attributes' paired comparisons, and the adaptive nature of the ASE provide an improvement over the traditional rating-based SEM approach.

We found that the predictive validity of the ACA and FPM was somewhat lower than that of the SEM stage of these methods (paired samples test not statistically significant, $p > .75$).[7] In other words, adding the 21 paired comparison questions in these methods did not help improve the predictive validity. Although this result may seem surprising, it is consistent with Srinivasan and Park's (1997) findings. Consistent with Toubia and colleagues (2003), we also found that the predictive validity of the FPM is somewhat greater (though not statistically significantly so) than that of the ACA. It should be noted that all preference measurement methods tested (ASE, ACA, FPM, and SEM) predicted significantly better ($p < .05$) than random choice.

Hierarchical Bayes estimation. In a series of analyses using multiple data sets, Sawtooth Software (2006) finds that estimating the ACA partworths using a hierarchical Bayes procedure can lead to improvement of 3%–12% in prediction ability. Similarly, Toubia and colleagues (2003) demonstrate an improvement of 3%–7% in predictive validity when estimating the FPM approach using a hierarchical Bayes procedure. To provide a fair comparison *among* the ASE, ACA, and FPM, we reestimated the partworths in these methods using a hierarchical Bayes estimation. Similar to the hierarchical Bayes ACA (denoted by ACA/HB) and hierarchical Bayes FPM (denoted by FPM/HB), in hierarchical Bayes ASE (denoted by ASE/HB), only the final regression is estimated using a hierarchical Bayes procedure from the data collected previously. As in the ASE, importances for attributes in intervals that were not opened (attributes for which no paired comparison questions were asked) were linearly interpolated. We estimated the ACA/HB and FPM/HB using Sawtooth Software's ACA/HB V. 2.0 software.[8] The ASE/HB was coded in GAUSS. In all methods, we used the first 15,000 iterations as a "burn-in" and the last 5000 iterations to estimate the conditional posterior distributions. Because the SEM does not involve any estimation, there is no corresponding SEM/HB.

Consistent with prior research, we find marginal improvement in predicting the choice set hit rates for estimating the partworths using the hierarchical Bayes approach for all

[7]When comparing the self-explicated stage of the ACA or FPM with the full methods and when comparing the classic and hierarchical Bayes estimation described subsequently, we conducted pairwise t-tests to compare the within-subject estimates, under the null hypothesis of no difference between the methods.

[8]We used Sawtooth's ACA/HB default setting for enforcing the self-explicated constraints both within and across attributes and using only the paired comparisons data for the hierarchical Bayes calibration.

three methods (6%, 3%, and 8% improvement for the ASE, ACA, and FPM, respectively; see Table 4). None of the improvements are statistically significant (paired samples test, $p > .4$). The prediction ability of the ASE/HB is substantially and significantly greater than that of the corresponding ACA/HB and FPM/HB. Thus, the improvement in predictive ability from using the hierarchical Bayes approach is small relative to the improvement achieved from using the ASE data collection approach over the existing data collection methods.

Aggregate-level predictions. Preferences are often measured to predict aggregate choice shares. Aggregate choice shares are used in conjoint simulators to guide managerial decisions; thus, significant differences between the methods in aggregate choice share predictions may have direct implications for product design or redesign decisions. To compare the alternative methods in terms of their ability to predict aggregate choice shares, we calculated the aggregate choice shares for each one of the four choice sets by aggregating respondents' highest-ranked alternatives in each choice set. We computed the mean absolute deviation (MAD) between the predicted and the actual choice shares for the four alternatives in each choice set and averaged the MAD across the four choice sets to obtain a measure of predictive validity.

In Table 5, we report the MAD of the alternative methods estimated using the individual estimates and the hierarchical Bayes approach. Consistent with the individual-level predictions, the ASE and ASE/HB had lower MAD (better predictive validity) than the alternative methods. Furthermore, the SEM predicted better than the ACA and FPM, and the FPM predicted better than the ACA. Unlike the individual-level predictions, the hierarchical Bayes procedure, which pools information across respondents, did not improve the aggregate-level predictions.

Overall, the individual and aggregate-level predictive analyses, using both classic and Bayesian estimation,

Table 5

COMPARISON OF AGGREGATE-LEVEL PREDICTIVE VALIDITY (MAD) (EMPIRICAL APPLICATION 1)

Method	*N*	*Individual Estimates*	*Hierarchical Bayes*
ASE	104	.067	.068
ACA	98	.113	.122
FPM	100	.096	.090
SEM	198	.082	—

Notes: Lower MAD indicates better predictive ability.

demonstrate that for the current empirical application, the predictive ability of the ASE is substantially and significantly better than that of the traditional SEM and two alternative adaptive methods, ACA and FPM. An important insight from Tables 4 and 5 is that the differences in data collection methods (i.e., SEM, ACA, FPM, and ASE) have a much greater impact on predictive validity than alternative statistical procedures (i.e., individual-level estimation versus hierarchical Bayes estimation).

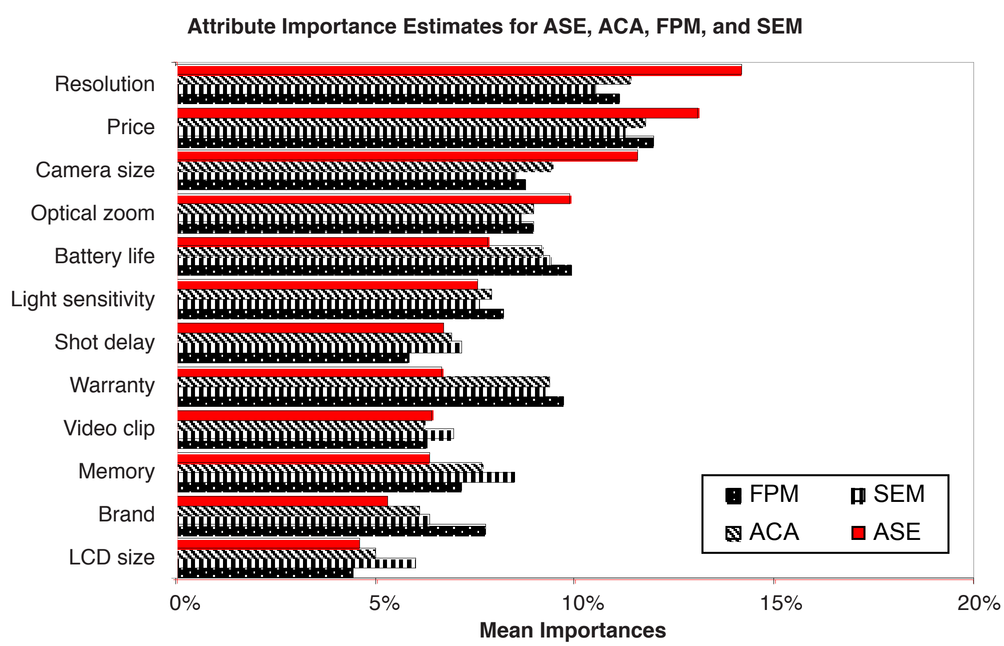
Estimated Importances

A possible reason for the improved predictions of the ASE over the alternative methods is that the ASE uses ranking and ratio-scaled trade-off questions to elicit the attribute importances rather than ratings scales used in the self-explicated stage and paired comparison of partial profiles in ACA and FPM. Prior research suggests that trade-offs induced by the ranking and constant-sum questions are likely to produce more variation in the attribute importances than the variations induced by rating scales (Krosnick and Alwin 1988).

In Figure 7, we plot the attribute importance distributions for the alternative preference measurement methods. The

Figure 7

COMPARISON OF AVERAGE ATTRIBUTE IMPORTANCE ESTIMATES (EMPIRICAL APPLICATION 1)



average attribute importance distribution of the ASE is indeed wider than those of the alternative methods. To compare the attribute importance distributions across methods, we calculated the coefficient of variation for each method. The coefficient of variation is the ratio of standard deviation divided by the mean of attribute importance estimates computed at the individual respondent level and averaged across respondents. The coefficient of variation of the ASE is significantly larger (almost double) than the coefficient of variation of all the alternative methods ($p < .01$; see Table 6). The hierarchical Bayes procedure, which tends to shrink the individual-level estimates toward the population mean, yields lower coefficients of variation. The coefficients of variation of the ACA and FPM are significantly larger than that of the SEM (paired samples test $p < .01$), suggesting that the paired comparison questions helped increase the variation produced by these methods' rating-based self-explicated stage.

The finding that the coefficient of variation of the ASE is much larger than that of the alternative methods, together with the improved predictive validity of the ASE, suggests that the ACA, FPM, and SEM underestimate the variation in importance ratings. To test for the "optimal" variation in the importance distributions, we varied the individual-level distribution of the estimated importances derived from the ASE using the following power transformation $I'_j = I_j^c / \Sigma_j I_j^c$, where I_j is the original ASE estimated importance for attribute j and $c > 0$ is a power parameter. For $c < 1$, the "new" variation in importances is narrower than the variation of the original estimated importances, whereas for $c > 1$, the modified importance distribution is wider than the original importance distribution. The hit rates measures are maximized for c between .7 and .8, suggesting that the ASE somewhat overestimates the variation of attribute importances. When we use the hierarchical Bayes approach to estimate the ASE attribute importances, the choice set hit rates are maximized for the original estimated importance ratings ($c = 1$). The optimal power parameters for the ACA, FPM, and SEM are 5.5, 4.0, and 1.0, respectively. However, even at their "optimal" power parameters, the predictive ability of the alternative methods was significantly poorer than the hit rates of the ASE at the original importance estimates (for details of this analysis, see the Web Appendix at http://www.marketingpower.com/jmrfeb11).

Reducing Respondents' Burden

One of the main advantages of adaptive methods such as the ASE, ACA, and FPM is the opportunity to reduce respondents' burden by asking fewer questions. Reducing respondents' cognitive load is particularly important for Web-based environments, for which respondents' patience tends to be low (Deutskens et al. 2004). Because in the adaptive algorithm we open the intervals in a decreasing order of uncertainty about the importance of attributes, it is likely that the marginal contribution of each additional paired comparison question is decreasing.

Consequently, we now test the effect of reducing the number of paired comparison questions on the predictive validity of the ASE. To do so, we estimated the individual partworths and corresponding predictive validity measures using only the first k paired comparison questions for each respondent. If the respondent has less than k paired comparison questions overall, we used the maximal number of paired comparisons asked for that respondent.

As Table 7 indicates, the ASE performs well even with just a few attribute importance paired comparison questions. With only five to seven paired comparison questions, the predictive validity of the ASE is similar to the predictive validity of the full ASE with all 21 paired comparison questions. The predictive validity of the ASE is maximal for 13 paired comparison questions, suggesting that beyond 13 paired comparison questions, we may be obtaining lower-quality data (respondent wear-out or fatigue). This result suggests that in running the ASE, α in Equation 3 should be set to a positive number. We further investigate this issue.

One of the major differences between the ASE and the alternative preference measurement methods is that the ASE uses a ranking task to measure attribute importance, as opposed to the rating-based scales the ACA, FPM, and SEM use. Previous research has found that ranking data has superior reliability (e.g., Miethe 1985; Munson and McIntyre 1979; Reynolds and Jolly 1980) and validity (e.g., Alwin and Krosnick 1985; Krosnick 1999) to rating data. The predictive validity of the ASE with a desirability ratings task, attribute rankings, and only one paired comparison question (comparing the most important attribute with the least important attribute to scale the rankings) already outperforms the ACA and the FPM with 21 paired comparison questions each and the SEM (the first row in Table 7). This result suggests that the ranking task plays an important role not only in facilitating the adaptive algorithm but also in directly improving the predictive ability of the ASE. In the current empirical application, the attribute ranking stage

Table 6

COMPARISON OF THE COEFFICIENT OF VARIATION OF THE IMPORTANCES (EMPIRICAL APPLICATION 1)

		Individual Estimates		*Hierarchical Bayes*	
Method	*N*	*M*	*SE*	*M*	*SE*
ASE	52	.849	.041	.728	.033
ACA	49	.487*	.016	.430*	.010
FPM	50	.515*	.018	.432*	.010
SEM	99	.403*	.014	—	—

*Statistically significantly lower than the corresponding ASE or ASE/HB ($p < .05$).

Table 7

PREDICTIVE VALIDITY MEASURES BY NUMBER OF ASE PAIRED COMPARISON QUESTIONS (EMPIRICAL APPLICATION 1)

Number of Questions (k)	*Choice Set Hit Rates*	*Pairs Hit Rates*	*Rank-Order Correlation*
1	.519	.670	.433
3	.577	.702	.478
5	.606	.696	.450
7	.606	.713	.488
9	.615	.729	.517
11	.635	.722	.508
13	.644	.734	.525
15	.625	.726	.517
17	.606	.716	.515
19	.625	.721	.523
21	.606	.718	.517

contributed approximately half the improvement in predictive validity over the competing methods, and the adaptive paired comparison questions contributed the other half. (The average hit rate of ACA, FPM, and SEM is .429. The choice set hit rate for ASE with only one paired comparison is .519, and the hit rate for the ASE method with all paired comparisons is .606.) We also tested the predictive ability of a naive approach that derives attribute importance from the range of the desirability ratings. This approach produced low predictive ability (for details of this analysis, see the Web Appendix at http://www.marketingpower.com/jmrfeb11).

In Table 8, we compare the predictive validity of the ACA, FPM, and ASE, estimated using only the first 11 paired comparison questions in each method. After only 11 paired comparison questions, the ASE and ASE/HB predict significantly better than the alternative methods. The difference in prediction ability between the ASE and the alternative methods (except the FPM/HB) is larger than the difference we report for the full 21 paired comparison questions.

Another approach for reducing the number of paired comparison questions in the ASE is to set a more restrictive termination cutoff parameter (α) in Equation 3. An advantage of this approach is that it allows for heterogeneity in the number of questions asked for each participant so that respondents with greater error rates (larger standard errors in their importance estimates) are terminated earlier. In Table 9, we outline the predictive validity of the ASE with varying termination cutoff parameter values. The highest predictive validity is obtained for α between .6 and .8, which corresponds to an average of 14–15 paired comparison questions. Consistent with the results in Table 7, we find that with a restrictive α of 2 (average of 5–7 paired comparison questions), we obtain similar predictive validity to the one with $\alpha = 0$. Further research should investigate more generally the appropriate value of α.

The preceding analyses emphasize that the ASE not only significantly improves predictive ability over existing preference measurement methods but also permits doing so with a relatively short questionnaire. A survey that includes desirability ratings, ranking, and 5–7 adaptive paired comparison questions already demonstrates much better predictive validity than the alternative methods.

The ASE log-linear regression. The log-linear regression used to estimate the attribute importances produced good fit for the constant-sum paired comparison data. The average adjusted R-square across respondents is .96. This high adjusted R-square suggests that treating the respondents' input to the constant-sum questions as ratio scaled is justified. Furthermore, the stated importance rankings from the ranking stage and the estimated importance ranking from the full ASE procedure were highly correlated (average rank order correlation $\rho = .82$). On average, only 1.9 paired comparison questions (11%) were inconsistent with respondents' original importance ranking; most of these are closer to the end of the survey when the "true" difference between importances is likely to be low. Additional analyses demonstrated low autocorrelation between questions that involve the same attribute and a low degree of heteroskedasticity in the log-linear regression. (For a full description of these analyses, see the Web Appendix at http://www.marketingpower.com/jmrfeb11.)

Postsurvey feedback. After the ASE, ACA, and FPM surveys, we asked respondents for feedback about their experience with the preference measurement task. Specifically, we asked them to rate on a seven-point scale the difficulty and clarity of the task, the degree of enjoyment derived from completing the task, and their personal assessment of how well the survey was able to capture their preferences. (This is only a perceived assessment of the ability of the method to capture preferences because respondents did not see their estimated preferences.) Respondents enjoyed the ASE significantly more than the ACA and FPM tasks ($p = .05$). However, respondents found the ACA task significantly clearer than the ASE ($p < .03$), but this difference was not significant between FPM and ASE, even though the FPM and ACA share identical respondent interfaces. There was no significant difference among the three methods in terms of the perceived task difficulty.

Survey duration. On average, the ASE took 906 seconds to complete, compared with 867 seconds for the ACA and 1296 seconds for the FPM. The difference between the ACA and the ASE is not statistically significant ($p > .1$), but the FPM took significantly longer than the ACA and ASE ($p < .01$), most likely because of the six- to eight-second delay between the paired comparison questions that possibly resulted from using interpreted code for the adaptive calculation of the best polyhedron. As we discussed previously, the ASE's performance with only seven paired comparison questions was similar to the ASE's performance with the full set of an average of 17.8 paired comparison questions. With only seven paired comparison questions, the ASE would have taken approximately 760 seconds to complete, making it the shortest survey. Empirical Application 1 provides strong support for the improved performance for the

Table 8

PREDICTIVE VALIDITY OF THE ASE, ACA, AND FPM WITH ONLY 11 PAIRED COMPARISON QUESTIONS (EMPIRICAL APPLICATION 1)

	Choice Set Hit Rates		*Pairs Hit Rates*		*Rank-Order Correlation*		
Method	*M*	*SE*	*M*	*SE*	*M*	*SE*	*Aggregate Choice Shares MAD*
ASE	.635	.052	.722	.025	.508	.056	.057
ACA	.398*	.044	.639*	.024	.333*	.057	.112
FPM	.390*	.054	.653*	.023	.382*	.052	.063
ASE/HB	.625	.051	.721	.025	.504	.056	.048
ACA/HB	.408*	.048	.670*	.020	.388*	.046	.086
FPM/HB	.490*	.086	.693	.036	.458	.051	.085

*Statistically significantly lower than the corresponding ASE or ASE/HB results ($p < .05$).

Notes: Lower MAD indicates better predictive validity.

Table 9

ASE'S PREDICTIVE VALIDITY MEASURES FOR DIFFERENT TERMINATION CUTOFF PARAMETER (α) VALUES (EMPIRICAL APPLICATION 1)

α	*Choice Set Hit Rates*	*Pairs Hit Rates*	*Rank-Order Correlation*	*Average Number of Paired Comparison Questions*
.0	.606	.718	.517	17.8
.2	.615	.718	.515	17.3
.4	.625	.720	.513	16.4
.6	.635	.726	.519	15.1
.8	.625	.721	.510	14.0
1.0	.625	.718	.498	12.2
1.2	.606	.716	.492	11.1
1.5	.577	.715	.504	9.3
2.0	.596	.712	.481	7.7
2.5	.606	.708	.487	6.5
3.0	.606	.707	.488	5.7

ASE approach relative to several competing models (ACA, FPM, and SEM).

EMPIRICAL APPLICATION 2: REPLICATION AND EXTENSIONS: LAPTOP COMPUTERS WITH 14 ATTRIBUTES

In the second empirical application involving laptop computers varying along 14 attributes, we compared the ASE with the ACA to replicate the results of the first empirical application. Furthermore, to disentangle the improvement obtained from breaking down the constant-sum attribute importance question across all attributes to a sequence of constant-sum paired comparison questions and the improvement offered by our adaptive design, we compared the ASE with an orthogonal fractional factorial design (hereinafter FFD), which includes a set of constant-sum paired comparison questions chosen from a fractional factorial design. (Therefore, there are no ranking of the attributes and no adaptive choice of paired comparison questions.)

In addition to the different product category, the increased number of attributes, and the comparison with the FFD, we also modified our validation procedure relative to Empirical Application 1. In this study, we included both pre- and post-preference measurement validation tasks. Thus, each respondent ranked four different choice sets, two before and two after the preference measurement task. Furthermore, in Empirical Application 1, we carefully chose four possible choice sets from an orthogonal design to maximize choice conflict. This enabled us to calculate aggregate prediction measures. However, such a validation task may be susceptible to idiosyncrasies of the chosen four choice sets. Therefore, in Empirical Application 2, we randomly chose the validation choice sets for each respondent from an orthogonal balanced design of 800 questions. To enhance choice conflict, we excluded 29 choice sets that had one or more of the following: five or more ties of attribute levels across any pair of alternatives in the set, an identical level for any of the attributes across the four alternatives, and strictly dominating alternatives. The average Kendall tau distance for the validation choice sets is .237. To allow for appropriate comparison across methods, we used the same set of randomly drawn validation questions in all three conditions (ASE, ACA, and FFD).

Similar to the digital cameras in Empirical Application 1, the domain of laptop computers is an appropriate product category for our purpose because of the large number of features often used to describe laptop computers and the relevance of this product category to our respondents (mostly students). On the basis of a pretest, we chose a set of 14 attributes and corresponding levels that respondents found most important (for a list of the 14 attributes and the corresponding levels, see Table 10).

Research Design

We randomly assigned 193 respondents, recruited from the behavioral lab of a large East Coast university, into three conditions: ASE (n = 65), ACA (n = 66), and FFD (n = 58).[9] The three groups were not statistically significantly different in terms of age, gender, laptop computer ownership, laptop computer familiarity, and degree of tech savviness ($p >$ .1). In all the conditions, respondents first completed a validation task followed by an unrelated filler task. Then respondents completed the corresponding preference measurement task, a postsurvey evaluation similar to the one used in the first empirical application, and a second validation task.

The ASE task was similar to the task described in Empirical Application 1, but with 14 attributes and up to 25 paired comparison questions. The average number of paired comparison questions asked per participant was 19.3, with a maximum of 25 and a minimum of 11. To make the ranking task easier for respondents, we first asked them to choose the seven attributes that are most important to them and rank these attributes in terms of their importance to them. Then,

[9]We omitted four respondents because of incomplete survey data or extremely short survey response time. Because Apple computers include a unique set of attributes and consumer preferences, we excluded participants who own an Apple laptop from our study.

Table 10

ATTRIBUTES AND LEVELS FOR THE LAPTOP COMPUTER STUDY (EMPIRICAL APPLICATION 2)

	Attribute	*Levels*
1.	Brand	Dell, Hewlett-Packard, Lenovo, Sony, Toshiba
2.	Battery life	2, 3.5, 5 hours
3.	Built-in Web camera	Not included, included
4.	CD/ DVD	Reads but does not burn DVDs and CDs, reads and burns DVDs and CDs
5.	Computer weight	3, 5, 7 lbs.
6.	Computer width	Ultra thin (1 in.), regular (1.5 in.)
7.	Hard drive	80, 120, 160, 240 GB
8.	Included software	Windows XP, Windows Vista, Windows XP + Office 2007, Windows Vista + Office 2007
9.	Memory (RAM)	1, 2, 4 GB
10.	Price	\$750, \$1,000, \$1,250, \$1,500
11.	Processor speed	1.5 GHz, 2 GHz, 2.5 GHz
12.	Screen resolution	XGA (1024 × 768), Super XGA (1280 × 1024), Ultra XGA (1600 × 1200)
13.	Screen size	11, 13, 15, 17 inches
14.	Warranty	No warranty or 1-, 2-, or 3-year warranty

in a subsequent screen, respondents ranked the remaining (less important) seven attributes. The ACA task included a self-explicated stage followed by 25 partial profile paired comparison questions. The first 13 paired comparison questions included profiles varying along two attributes, and the last 12 questions included profiles varying along three attributes.

In the FFD, respondents first completed a desirability rating task, similar to the one used in the ASE and ACA, followed by 21 constant-sum paired comparison questions. The interface of the constant-sum paired comparison questions looked similar to that of the ASE. We chose a subset of 21 questions from all possible paired comparisons using an orthogonal balanced incomplete block design (Clatworthy 1973). This approach guarantees that each product attribute will appear in three paired comparison questions (21 pairs × 2 attributes per pair/14 attributes; for a related approach, see Johnson and VanDyk 1975). The order of pairs was randomized across respondents within three blocks of seven suborthogonal designs.

Predictive Validity

Table 11 reports the predictive validity measures for the ASE, ACA, and FFD. Because we did not find a significant order effect between the pre- and postpreference measurement validation tasks and between the two choice sets within each validation task, we aggregated the results across the four validation choice sets for each respondent. Similar to Empirical Application 1, we use hit rates, pairs hit rates, and correlations to compare the alternative methods.

The ASE correctly predicted the highest-ranked alternatives in 54% of the choice sets and the pairs hit rates in 71% of the pairs. The results of the laptop study replicated the results of Empirical Application 1 in terms of the superior predictive ability of the ASE relative to the ACA, though the magnitude of the improvement is somewhat lower. The ASE provided 17% improvement in validation choice set hit rates over the ACA.[10] The improvement in predictive ability of the ASE relative to the ACA is statistically significant ($p <$ .05) for the choice set hit rates and rank-order correlation and marginally significant ($p = .07$) for the pairs hit rates.

[10]Although we report only the average predictive ability measures across the pre- and postpreference measurement task validations, the ASE outpredicted the ACA and the FFD on all three measures also in the pre- and postpreference measurement task validations separately.

Table 11

COMPARISON OF INDIVIDUAL-LEVEL PREDICTIVE VALIDITY (EMPIRICAL APPLICATION 2)

	Choice Set Hit Rates		Pairs Hit Rates		Rank-Order Correlation	
Method (Sample Size)	*M*	*SE*	*M*	*SE*	*M*	*SE*
ASE (n = 65)	.542	.033	.712	.017	.498	.037
ACA (n = 66)	.462*	.032	.680	.014	.423*	.030
FFD (n = 58)	.461*	.036	.693	.015	.447	.036
ASE/HB (n = 65)	.554	.033	.716	.017	.498	.037
ACA/HB (n = 66)	.511	.031	.692	.012	.451	.027
FFD/HB (n = 58)	.474*	.036	.699	.016	.464	.036

*Statistically significantly lower than the corresponding ASE or ASE/HB results ($p < .05$).

The ASE also predicted the validation choices significantly better than the FFD method ($p < .05$) and directionally better for the pairs hit rates and rank-order correlations. This suggests that the adaptive choice of the constant-sum paired comparison questions played an important role in the improved performance of the ASE beyond the improvement obtained from merely breaking down the constant-sum attribute importance question across all attributes into a set of constant-sum paired comparison questions. Because the ASE method involves adaptive choice of questions, it is potentially susceptible to endogeneity bias (Hauser and Toubia 2005). The improved predictive ability of ASE relative to an FFD method (which, by definition, does not suffer from endogeneity bias) suggests that to the extent endogeneity exists in the ASE, it does not make predictions worse. Across the three prediction measures, the FFD predicted the validation tasks choices better than the ACA; however, this difference is not statistically significant. This result is consistent with the result of a prior study (available on request) that compared the FFD with the ACA using MBA students' actual job choice preferences. We also compared the alternative methods using only the first 14 paired comparison questions. Similar to the results of Empirical Application 1, the prediction ability of the ASE was not hurt by the shorter questionnaire and was superior to that of the ACA. (The full set of results can be obtained on request.)

We compared the predictive ability of the three methods when the attribute importances and partworths were estimated using the hierarchical Bayes approach (see Table 11). In all methods, we used the first 40,000 iterations as a "burn-in" and the last 10,000 iterations to estimate the conditional posterior distributions. Consistent with the first empirical application and prior literature, for all three methods the hierarchical Bayes estimation improved the prediction ability relative to the individual-level estimation, though none of these improvements are statistically significant (paired samples test, $p > .3$). Moreover, in all three measures, the prediction ability of the ASE/HB is superior to that of the ACA/HB and the FFD/HB. However, unlike Empirical Application 1, the improved prediction ability of the ASE/HB relative to the ACA/HB did not reach statistical significance ($p = .17$). The ASE/HB predicted the choice set hit rates significantly better than the FFD/HB ($p = .05$).

Overall, in comparing the ASE and ACA, the results of Empirical Application 2 were consistent with the results of Empirical Application 1, though smaller in magnitude. The comparison of the ASE with the FFD highlights that the adaptive aspect of the ASE is an important component of the proposed approach.

SUMMARY AND CONCLUSIONS

In this research, we propose an improved self-explicated approach for multiattribute preference measurement (conjoint analysis) when the number of attributes is large (i.e. ten or more). We view the contribution of the proposed approach over existing approaches as threefold. First, we break down the constant-sum (across all attributes) importance questions into ranking of attributes together with a set of constant-sum paired comparison attribute importance questions. Second, because of the ratio-scale nature of attribute importance ratings and the constant-sum question format, our proposed constant-sum paired comparison questions are an inherently different format from the rating-based partial profile paired comparison questions used in ACA and FPM. Third, our adaptive design chooses the constant-sum paired comparison questions to maximize the information elicited from each question. Specifically, questions about the larger attribute importance differences are asked first, thus recovering much of the required information with relatively few questions. The adaptive design reduces respondents' burden and makes it possible to increase the number of attributes studied. In two empirical applications using digital cameras and laptop computers with 12 and 14 attributes, respectively, we find that the percentage of top choices correctly predicted by the ASE is significantly and substantially higher than that of the ACA, FPM, and the rating-based SEM. The ASE approach leads to wider importance distributions relative to the ACA, FPM, and SEM. Furthermore, we find that even for a relatively short questionnaire with desirability ratings questions, attribute importance ranking , and only five to seven paired comparison questions, the ASE produced better predictions than the alternative methods. Another advantage of the proposed ASE is that it uses a simple log-linear regression to estimate the derived attribute importances, making it easy to implement. A key insight from our empirical results is that alternative data collection methods (i.e., SEM, ACA, FPM, and ASE) have a much greater impact on predictive validity than variations in statistical procedures (i.e., individual-level estimation versus hierarchical Bayes).

In Empirical Application 1, the attributes' ranking stage in the ASE contributed to approximately half the improvement in predictive validity of the ASE relative to the competing methods, and the adaptive paired comparison questions contributed the other half. In Empirical Application 2, we further examined this issue by comparing the ASE with an approach that breaks down the attribute importance constant-sum question to a set of attribute importance constant-sum paired comparison questions chosen from an orthogonal design FFD. The improved performance of the ASE relative to the FFD suggests that the ranking stage and the adaptive choice of the constant-sum paired comparison questions significantly contribute to attribute importance elicitation. The directional improved predictive ability of the FFD relative to the ACA suggests that the proposed questionnaire format of constant-sum paired comparison attribute importance questions (without the adaptive design) is at least as good as the adaptive partial profile paired comparison questions used in the ACA.

In recent years, partial profile choice-based conjoint analysis techniques have been gaining popularity. However, both academics and practitioners are skeptical about the ability of these methods to estimate stable individual-level partworths for a large number of attributes even with a hierarchical Bayes approach (Orme 2007). Further research could compare the ASE with the partial profile choice-based conjoint analysis for a large number of attributes.

Recent studies have demonstrated the potential in applying incentive-compatible or incentive-aligned mechanisms in conjoint analysis studies (Ding 2007; Ding, Grewal, and Liechty 2005). Although the application of incentive alignment to ASE is not as straightforward as in choice-based conjoint analysis, in which respondents have a chance to win one of the profiles they chose during the preference measurement task, we believe that incentive alignment will benefit the ASE as well.

The proposed approach could span beyond measuring attribute weights in conjoint analysis–like settings to more generally estimating the priorities people attach for a long list of items. We have some preliminary evidence that the ASE method could be successfully applied to prioritizing research topics and voter issues.

It would be useful to replicate our empirical application with larger sample sizes and using additional product and service categories. In addition, further research could examine the performance of the ASE for problems with an even larger number of attributes.

REFERENCES

Agarwal, Manoj K. and Paul E. Green (1991), "Adaptive Conjoint Analysis Versus Self-Explicated Models: Some Empirical Evidence," *International Journal of Research in Marketing*, 8 (2), 141–46.

Alwin, Duane F. and Jon A. Krosnick (1985), "The Measurement of Values in Surveys: A Comparison of Ratings and Rankings," *Public Opinion Quarterly*, 49 (4), 535–52.

Axelrod, Joel N. (1968), "Attitude Measures that Predict Purchase," *Journal of Advertising Research*, 8 (March), 3–18.

Bradlow, Eric T. (2005), "Current Issues and a 'Wish List' for Conjoint Analysis," *Applied Stochastic Models in Business and Industry*, 21 (4–5), 319–23.

Carmone, Frank J., Ali Kara, and Stelios H. Zanakis (1997), "A Monte Carlo Investigation of Incomplete Pairwise Comparison," *European Journal of Operational Research*, 102 (3), 538–53.

Clatworthy, Willard H. (1973), *Tables of Two Associates Partially Balanced Designs: National Bureau of Standards, Applied Mathematics Series 63*. Washington, DC: U.S. Department of Commerce.

Cui, Dapeng and David Curry (2005), "Prediction in Marketing Using the Support Vector Machine," *Marketing Science*, 24 (4), 595–615.

Deutskens, Elisabeth, Ko De Ruyter, Martin Wetzels, and Paul Oosterveld (2004), "Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study," *Marketing Letters*, 15 (1), 21–36.

Ding, Min (2007), "An Incentive-Aligned Mechanism for Conjoint Analysis," *Journal of Marketing Research*, 44 (May), 214–23.

———, Rajdeep Grewal, and John Liechty (2005), "Incentive-Aligned Conjoint Analysis," *Journal of Marketing Research*, 42 (February), 67–82.

Evgeniou, Theodoros, Constantinos Boussios, and Giorgos Zacharia (2005), "Generalized Robust Conjoint Estimation," *Marketing Science*, 24 (3), 415–29.

Green, Paul. E., Stephen M. Goldberg, and Mila Montemayor (1981), "A Hybrid Utility Estimation Model for Conjoint Analysis," *Journal of Marketing*, 45 (Winter), 33–41.

———, Abba M. Krieger, and Manoj K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research*, 28 (May), 215–22.

——— and Vithala R. Rao (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research*, 8 (August), 355–63.

——— and V. Srinivasan (1978), "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5 (September), 103–123.

——— and ——— (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing Research*, 54 (October), 3–19.

Harker, Patrick T. (1987), "Incomplete Pairwise Comparisons in the Analytic Hierarchy Process," *Mathematical Modelling*, 9 (11), 837–48.

Hauser, John R. and Olivier Toubia (2005), "The Impact of Utility Balance and Endogeneity in Conjoint Analysis," *Marketing Science*, 24 (3), 498–507.

———, ———, Theodoros Evgeniou, Daria Dzyabura, and Rene Befurt (2010), "Disjunctions of Conjunctions, Cognitive Simplicity, and Consideration Sets," *Journal of Marketing Research*, 47 (June), 485–96.

Huber, Joel, Dick R. Wittink, John A. Fiedler, and Richard Miller (1993), "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice," *Journal of Marketing Research*, 30 (February), 105–114.

Johnson, Richard M. (1987), "Adaptive Conjoint Analysis," in *Sawtooth Software Conference Proceedings*. Ketchum, ID: Sawtooth Software, 253–65.

——— and Gerald J. VanDyk (1975), "A Resistance Analogy for Efficiency of Paired Comparison Designs," working paper, Market Facts, Inc.

Krieger, Abba M., Paul E. Green, and Yoram J. Wind (2004), "Adventures in Conjoint Analysis: A Practitioner's Guide to Trade-Off Modeling and Applications," working paper, Wharton School, University of Pennsylvania.

Krosnick, Jon A. (1999), "Maximizing Questionnaire Quality," in *Measures of Political Attitudes*, John P. Robinson, Philip, R. Sharer, and Lois S. Wrightsman, eds. New York: Academic Press, 37–52.

——— and Duane F. Alwin (1988), "A Test of the Form-Resistant Correlation Hypothesis," *Public Opinion Quarterly*, 52 (Winter), 526–38.

Leigh, Thomas W., David B. MacKay, and John O. Summers (1984), "Reliability and Validity of Conjoint Analysis and Self-Explicated Weights: A Comparison," *Journal of Marketing Research*, 21 (November), 456–62.

Marder, Eric (1997), *The Laws of Choice: Predicting Customer Behavior*. New York: The Free Press.

Marshall, Pablo and Eric T. Bradlow (2002), "A Unified Approach to Conjoint Analysis Models," *Journal of the American Statistical Association*, 97 (459), 674–82.

Miethe, Terance D. (1985), "The Validity and Reliability of Value Measurements," *Journal of Psychology*, 119 (5), 441–53.

Munson, Michael J. and Shelby H. McIntyre (1979), "Developing Practical Procedures for the Measurement of Personal Values in Cross-Cultural Marketing," *Journal of Marketing Research*, 16 (May), 48–52.

Netzer, Oded, Olivier Toubia, Eric T. Bradlow, Ely Dahan, Theodoros Evgeniou, Fred M. Feinberg, et al. (2008), "Beyond Conjoint Analysis: Advances in Preference Measurement," *Marketing Letters*, 19 (3), 337–54.

Orme, Bryan K. (2007), "Which Conjoint Method Should I Use?" research paper series, Sawtooth Software.

Park, Young Hoon, Min Ding, and Vithala R. Rao (2008), "Eliciting Preference for Complex Products: A Web-Based Upgrading Method," *Journal of Marketing Research*, 45 (October), 562–74.

Rao, Vithala R. and John R. Hauser (2004), "Conjoint Analysis, Related Modeling, and Application," in *Market Research and Modeling: Progress and Prospects: A Tribute to Paul Green*, Y. Wind and P.E. Green, eds. New York: Kluwer Academic Publishers, 141–68.

Reynolds, Thomas J. and James P. Jolly (1980), "Measuring Personal Values: An Evaluation of Alternative Methods," *Journal of Marketing Research*, 17 (November), 531–36.

Saaty, Thomas L. (1980), *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. New York: McGraw-Hill.

Sawtooth Software (2006), "The ACA/Hierarchical Bayes v3.0 Technical Paper," technical paper series, Sawtooth Software.

Scholz, Sören W., Martin Meissner, and Reinhold Decker (2010), "Measuring Consumer Preferences for Complex Products: A Compositional Approach Based on Paired Comparisons," *Journal of Marketing Research*, 47 (August), 685–98.

Silk, Alvin J. and Glen L. Urban (1978), "Pre-Test-Market Evaluation of New Packaged Goods: A Model and Measurement Methodology," *Journal of Marketing Research*, 15 (May), 171–91.

Srinivasan, V. (1988), "A Conjunctive-Compensatory Approach to the Self-Explication of Multiattributed Preferences," *Decision Sciences*, 19 (Spring), 295–305.

——— and Chan Su Park (1997), "Surprising Robustness of the Self-Explicated Approach to Customer Preference Structure Measurement," *Journal of Marketing Research*, 34 (May), 286–91.

——— and Gordon A. Wyner (1989), "CASEMAP: Computer-Assisted Self-Explication of Multi-Attributed Preferences," in *New Product Development and Testing*, W. Henry, M. Menasco, and H. Takada, eds. Lexington, MA: Lexington Books, 91–111.

Ter Hofstede, Frenkel, Youngchan Kim, and Michel Wedel (2002), "Bayesian Prediction in Hybrid Conjoint Analysis," *Journal of Marketing Research*, 36 (May), 253–61.

Toubia, Olivier, John R. Hauser, and Rosanna Garcia (2007), "Probabilistic Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis: Theory and Application," *Marketing Science*, 26 (5), 596–610.

———, ———, and Duncan I. Simester (2004), "Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis," *Journal of Marketing Research*, 41 (February), 116–31.

———, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, 22 (Summer), 273–303.

Weiss, Elliott N. and Vithala R. Rao (1987), "AHP Design Issues for Large Scale Systems," *Decision Sciences*, 18 (1), 43–61.

Wind, Jerry, Paul E. Green, Douglas Shifflet, and Marsha Scarbrough (1989), "Courtyard by Marriott: Designing a Hotel Facility with Consumer-Based Marketing Models," *Interfaces*, 19 (January–February), 25–47.