MARKETING SCIENCE
Vol. 24, No. 3, Summer 2005, pp. 498–507
ISSN 0732-2399 | EISSN 1526-548X | 05 | 2403 | 0498

informs

DOI 10.1287/mksc.1040.0108

The Impact of Utility Balance and Endogeneity in Conjoint Analysis

John R. Hauser
MIT Sloan School of Management, Massachusetts Institute of Technology, E56-314, 38 Memorial Drive, Cambridge, Massachusetts 02142, jhauser@mit.edu

Olivier Toubia
Columbia Business School, Columbia University, 522 Uris Hall, 3022 Broadway, New York, New York 10027, ot2107@columbia.edu

Adaptive metric utility balance is at the heart of one of the most widely used and studied methods for conjoint analysis. We use formal models, simulations, and empirical data to suggest that adaptive metric utility balance leads to partworth estimates that are relatively biased—smaller partworths are upwardly biased relative to larger partworths. Such relative biases could lead to erroneous managerial decisions. Metric utility-balanced questions are also more likely to be inefficient and, in one empirical example, contrary to popular wisdom, lead to response errors that are at least as large as nonadaptive orthogonal questions. We demonstrate that this bias is because of endogeneity caused by a "winner's curse." Shrinkage estimates do not mitigate these biases. Combined with adaptive metric utility balance, shrinkage estimates of heterogeneous partworths are biased downward relative to homogeneous partworths. Although biases can affect managerial decisions, our data suggest that, empirically, biases and inefficiencies are of the order of response errors. We examine viable alternatives to metric utility balance that researchers can use without biases or inefficiencies to retain the desired properties of (1) individual-level adaptation and (2) challenging questions.

Key words: conjoint analysis; efficient question design; adaptive question design; Internet market research; e-commerce; product development
History: This paper was received August 26, 2003, and was with the authors 8 months for 2 revisions; processed by Gary Lilien.

1. Motivation

Adaptive conjoint analysis (ACA) has been used widely, both academically and commercially, for more than 25 years. Many authors have studied and improved both the theory and practice of ACA and many firms have relied on ACA for both product development and advertising decisions (e.g., Allenby and Arora 1995, Carroll and Green 1995, Choi and DeSarbo 1994, Green and Krieger 1995, Green et al. 1991, Huber et al. 1993, Toubia et al. 2003). Sawtooth Software claims that ACA is one of the most popular forms of conjoint analysis in the world, likely has the largest installed base, and, in 2001, accounted for 37% of their sales (private communications and *Marketing News* April 1, 2002, p. 20). To its credit, Sawtooth Software has responded to academic critique with improvements through five generations. For example, hierarchical Bayes (HB) estimation is now available and ACA v.5 addresses the scaling issues highlighted by Green et al. (1991). However, while both scaling and estimation have improved steadily, question selection in ACA has not changed since "they were originally programmed for the Apple II computer in the late 70s (Orme and King 2002)."

Adaptive *metric* utility balance has always been at the heart of ACA question selection. That is, (1) the preference scale is metric (interval, not ordinal), (2) paired-comparison questions are chosen adaptively based on prior responses by individual respondents, and (3) the key criterion is utility balance (subject to other balance constraints). By utility balance, we mean that "ACA presents to the respondent pairs of concepts that are as nearly equal as possible in estimated utility" (Sawtooth Software 2002, p. 11).

In recent years, in an attempt to improve on the philosophy of ACA, researchers have begun to experiment with different forms of question adaptation. One set of researchers has explored "aggregate customization" for choice-based questions (Arora and Huber 2001; Huber and Zwerina 1996; Johnson et al. 2003; Kanninen 2002; Orme and Huber 2000; Sandor and Wedel 2001, 2002, 2003). These researchers retain the utility-balanced criterion as *one* criterion in their algorithms, but focus on choice questions (one profile

Table 1 Illustrative Example of Relative Bias Because of Adaptive Metric Utility Balance

Hypothetical features	"True" partworths	ACA questions	Percent difference (%)	Percent difference, normalized (%)	"True" willingness to pay ($)	Estimated willingness to pay ($)
Handle	10	11.8	18.1	5.6	15	15
Price	20	23.6	18.0	5.6	30	30
Logo	30	34.0	13.3	1.4	45	43
Closure	40	45.0	12.4	0.6	60	57
Mesh pocket	50	55.4	10.8	−0.8	75	70
PDA holder	60	66.7	11.2	−0.5	90	85
Cell phone	70	77.4	10.6	−1.0	105	98
Color	80	88.7	10.8	−0.8	120	113
Size	90	100.5	11.6	−0.1	135	128
Boot	100	111.6	11.6	−0.2	150	142

is chosen from a set) rather than metric questions. They adapt questions between respondents (pretest, then full-scale study) rather than within respondents.[1] Another set of researchers drop the utility-balanced criterion, but use metric questions that are adapted within respondents (Toubia et al. 2003).

This paper focuses on the impact of adaptive metric utility balance. We show that this criterion often leads to biases, inefficiencies, and, potentially, higher response errors. We provide examples where these biases and inefficiencies can adversely affect managerial decisions, but, fortunately, in most cases, the magnitude of the effects is modest. Nonetheless, the phenomena are real and can be easily avoided. In particular, the biases and inefficiencies can be mitigated with the use of choice questions and/or polyhedral methods.

2. An Illustrative Example

We draw on an application in which ACA was used as an aid to the design of a laptop computer bag with nine binary features plus price, specified at two levels—$100 and $70 (Toubia et al. 2003). We examine that data below, but first consider a hypothetical example in which all respondents are homogeneous and the true partworth differences are $10, 20, 30, \ldots, 100$ as shown in the second column of Table 1. For example, the partworth of "no handle" is -5 and the partworth of having a handle on the bag is $+5$. We simulate 1,000 respondents as follows:

- The a priori self-explicated questions (SEs) are chosen to be unbiased with normally distributed noise (ACA needs the SEs to select questions).
- Twenty metric paired-comparison questions are chosen by the utility-balance criterion using ACA's question-selection algorithm.
- Respondent answers are unbiased with normally distributed noise.
- Estimation uses standard ordinary least squares (OLS) estimation (later in this paper, we examine the impact of hierarchical Bayes estimation).

The results, shown in the third column of Table 1, suggest that the estimates based on ACA questions are upwardly biased and that the bias increases with the magnitude of the true partworth differences. The fourth column suggests that the bias increases *less than proportionally*—there is relative bias. Features with low partworths are biased proportionally more than features with high partworths. This relative bias survives normalization (column 5 of Table 1). Table 1 is illustrative—we can make the bias larger (or smaller) with other examples. The exact parameters for this simulation, and all simulations in this paper are available in an online appendix.

The data of Green et al. (1991) anticipate the upward bias in partworths, but not the relative bias. They hypothesize that the bias "results as subjects attempt to utilize the full range of the (metric) scale" (Green et al., p. 219). Such "stretching" bias is not in our simulations, thus the bias in Table 1 must be because of another effect. However, the Green-Krieger-Agarwal (GKA) effect would reinforce the bias identified in Table 1. We return to the GKA effect later in this paper.

The relative bias is modest, but it can affect managerial decisions. We compute "true" and estimated willingness to pay (WTP) in the last two columns of Table 1. For features with small partworths, the differences are barely noticeable, but for the partworths of important and costly features, the differences are larger. If the "boot" cost $145 to manufacture, the true partworths would imply it should be included, but the estimated partworths would not. Aggregated over the nine hypothetical features the estimated partworths underpredict WTP by approximately 5.6%. In some product categories, this could be a managerially significant percentage. For example, Colgate introduced body washes with a unique no-leak "Zeller

[1] For example, the title of the Huber and Zwerina (1996) paper is "The Importance of Utility Balance in Efficient Choice Designs." Many of the other papers build on that paper. The Johnson et al. (2003) paper adapts choice questions for each respondent based on prior self-explicated (SE) questions, utility balance, and efficiency.

valve" cap that enabled bottles to be stored cap-side down. The increase in cost was only a few percentage points. However, subsequent market research suggested that consumers' WTP for body washes did not justify this improvement (private communication). Had adaptive metric utility balance been used, even a small bias in estimated WTP would have caused Colgate to miss the substantial savings from dropping the Zeller valve.

Empirical data differ from the illustrative example in many ways. The SEs may not be unbiased, there may be heterogeneity in respondents' partworths, and the errors, both in the SEs and in the metric paired-comparison questions may be larger or smaller than in our simulation. Furthermore, most comparative empirical experiments are between groups of respondents rather than within respondents.

As an example, consider the empirical data from Toubia et al. (2003) in which 88 randomly assigned respondents answered fixed orthogonal questions and 80 randomly assigned respondents answered ACA-generated questions. The average OLS partworth estimates based on 16 questions are shown in Table 2. The estimated mean partworths in Table 2 are dramatically different. As in our illustrative example, the ACA estimates suggest different managerial decisions. The ACA estimates suggest that, on average, a handle will be bought at \$20, but the orthogonal questions suggest otherwise.

Notice that the percent differences in Table 2 are much larger than those in Table 1. On average, partworths are 42% larger when ACA questions are used than when orthogonal questions are used. In Table 1, the percent difference is negatively correlated with the true partworths ($r = -0.78$, $t = -3.5$). In Table 2, the correlation with the orthogonal partworths is negative, but not significant ($r = -0.23$, $t = -0.66$) and slightly smaller for WTP ($r = -0.14$, $t = -0.39$). (The correlations do not change if we use normalized partworths.) Not obtaining significance is not surprising because empirical data are less precise than "known" homogeneous partworths. In addition, the Table 2 comparison is between groups of respondents who might vary slightly in their true partworths because of finite sampling, there is heterogeneity within groups, and, even if the orthogonal question estimates are unbiased, they are subject to response errors. Furthermore, as discussed later, there might be other sources of empirical noise and/or bias in metric utility-balanced adaptive questions than the systematic bias highlighted in Table 1.

Table 1 is illustrative and Table 2 is only suggestive, but they anticipate our theoretical findings. Adaptive metric utility questions are relatively biased because of endogeneity. We also demonstrate that such questions are inefficient and may induce more response error. Although the effects are modest, they can affect managerial decisions. We also demonstrate that these biases and inefficiencies (and perhaps greater response error) can be avoided with other question-selection methods that are now available. Finally, we demonstrate why "shrinkage" estimates do not overcome these biases and may, themselves, introduce new biases when coupled with adaptive metric utility-balanced questions.

3. Endogeneity and Adaptive Metric Utility Balance

When questions are selected adaptively based on previous answers by a respondent, there is the potential for endogeneity bias because new questions might depend upon the errors made by respondents in their previous answers (Judge et al. 1985, p. 571). If such endogeneity bias is to produce the results in Table 1, it must be systematic (all partworths are biased upward) and relative (smaller partworths are biased relatively more than larger partworths). In this section, we explore whether adaptive utility-balanced *metric* questions cause systematic and relative endogeneity bias. We use a stylized model to understand and illustrate the cause of the biases, then provide a more general explanation based on the winner's

Table 2 Differences in Average Utility Weights Between ACA and Orthogonal Questions

Actual features	Orthogonal questions	ACA questions	Percent difference (%)	Percent difference, normalized (%)	Orthogonal WTP (\$)	ACA WTP (\$)
Handle	28.0	55.5	98	42	15.54	27.91
Price	54.0	59.6	10	−21	30.00	30.00
Logo	24.3	18.4	−24	−46	13.49	9.26
Closure	13.3	22.8	72	23	7.33	11.48
Mesh pocket	7.3	10.4	42	1	4.07	5.22
PDA holder	9.4	−2.6	−127	−119	5.21	−1.29
Cell phone	11.2	8.6	−23	−45	6.23	4.34
Color	27.7	49.8	80	28	15.39	25.03
Size	7.3	38.2	425	275	4.04	19.20
Boot	22.2	25.6	15	−18	12.33	12.87

curse. Finally, we provide simulations that isolate the bias as systematic winner's curse endogeneity.

Formal Analysis of a Simple Problem

Consider products with two binary features (with levels denoted by 0 and 1) and assume no interactions among the features. Scale the low level of each feature to zero and let w_i be the partworth of the high level of feature i. Denote the utility of a product with feature 1 and feature 2 by u(feature 1, feature 2). There are four possible profiles with true utilities given by

$$u(0,0)=0,$$
$$u(1,0)=w_1,$$
$$u(0,1)=w_2, \quad \text{and}$$
$$u(1,1)=w_1+w_2.$$

Assume response error is an additive, zero-mean random variable, ε, with probability distribution $f(\varepsilon)$. For example, if the respondent is asked to compare $\{1,0\}$ to $\{0,1\}$, the answer is given by $w_1-w_2+\varepsilon$. If he or she is asked to compare $\{1,1\}$ to $\{0,1\}$, the answer is given by $w_1+\varepsilon$. Denote the estimates of the partworths with $\widehat{w}_1$ and $\widehat{w}_2$.

Without loss of generality, assume the off-diagonal question, which compares $\{0,1\}$ to $\{1,0\}$, is the most utility-balanced first question and that $w_1>w_2$. Label the error associated with the first question as ε_{ub} and label errors associated with subsequent questions as either ε_1 or ε_2. In this simple problem, adaptive metric utility balance implies the following sequence:

$$\text{First question:} \quad \widehat{w}_1-\widehat{w}_2=w_1-w_2+\varepsilon_{ub},$$

$$\text{Second question:} \quad \widehat{w}_1=w_1+\varepsilon_1 \quad \text{if } \varepsilon_{ub}<w_2-w_1, \tag{Case 1}$$

$$\widehat{w}_2=w_2+\varepsilon_2 \quad \text{if } \varepsilon_{ub}\geq w_2-w_1. \tag{Case 2}$$

Because the second question depends upon the error in the first question, there is endogeneity. We demonstrate in an appendix (available from the authors) that Case 1 leads to upward bias in $\widehat{w}_2$ and Case 2 leads to upward bias in $\widehat{w}_1$. The intuitive idea is that the second question depends on the error in the respondent's answer to the first question. Overall bias depends upon the probabilities that each case is chosen, times the conditional expectations of the estimates for each case. The online appendix shows that biases are always positive and relative. (We show $\{(E[\widehat{w}_2]-w_2)/w_2\}-\{(E[\widehat{w}_1]-w_1)/w_1\}>0$.) This mathematical result applies formally to the 2×2 stylized model, but we feel it illustrates the basic phenomenon that applies more generally.

Proposition 1. *For a simple problem involving two binary features, adaptation based on metric utility balance* (1) *biases partworth estimates upward and* (2) *biases smaller partworths proportionally more than larger partworths*.

More General Analysis—The Winner's Curse

In the stylized model, the second conjoint question focuses on the partworth, which is believed to be smaller based on the answer to the first question. However, this belief is influenced by noise and may be inaccurate. This is the basic, generalizable source of the bias: when we select the question we *predict* to be most utility balanced, we overestimate (in expectation) its level of utility balance. More formally, we choose the next question from a pool of questions $(q_i)_{i\in I}$ and, based on the previous answers, we estimate the absolute value of the answer, $\widehat{w}_i$. However, if the estimate is a random variable with mean $\overline{w}_i$, the very act of choosing the question q_{i^*} with the smallest $\widehat{w}_i$ implies that $E(\widehat{w}_{i^*})<\overline{w}_{i^*}$. Indeed, the expectation of a random variable conditional on it being the smallest from a set of random variables is lower than its unconditional expectation. This is the same probabilistic phenomenon as the winner's curse, a well-known result in auction theory: the winner of a first-price auction for a common value good is "cursed" by the act of winning and pays too high a price (Capen et al. 1971, Kagel and Levin 1986, Thaler 1992).

The winner's curse is consistent with GKA's observation that respondents' answers use a larger range of the response scale (than predicted); however, the underlying mechanism is different. The winner's curse is because of endogenous question selection rather than a change in the respondents' reactions to the questions.

To see this formally, let $\vec{w}_q$ be the estimated vector of binary partworths, estimated after the qth paired-comparison question. We use binary features to simplify exposition, and without loss of generality, scale the low level of each feature to zero. The same arguments apply, but with more cumbersome notation, to multilevel features and to standard ACA scaling. Let $\vec{x}_q$ be the row vector corresponding to the qth question, and let X_q be the matrix of the first q questions obtained by stacking the q row vectors. If a_{q+1} is the answer to the $(q+1)$st question, then the predicted answer is $\hat{a}_{q+1}=\vec{x}_{q+1}\vec{w}_q$ and the $(q+1)$st estimate can be obtained from the qth estimate by the following equation (Sawtooth Software 2002, p. 19, Equation 5).[2]

$$\vec{w}_{q+1}-\vec{w}_q=(I+X_q'X_q)^{-1}\vec{x}_{q+1}'\cdot\frac{a_{q+1}-\hat{a}_{q+1}}{1+\vec{x}_{q+1}(I+X_q'X_q)^{-1}\vec{x}_{q+1}'}. \tag{1}$$

[2] ACA includes the SE questions when selecting the next question. For p parameters, this prefaces a p-dimensional identity matrix, I, to the top of X_q.

Because $I + X_q'X_q$ is positive definite, the denominator is positive. After balancing utility, ACA randomly selects one of the profiles as the right side of the comparison. Thus, in Equation (1), without loss of generality, we can orient questions mathematically such that the predicted answer $\hat{a}_{q+1}$ is nonnegative. (Actual empirical questions are randomly oriented.)

The winner's curse causes a_{q+1} to be larger than $\hat{a}_{q+1}$ in expectation. Let $\vec{e}$ be a row vector of 1's; then $\vec{e}(\vec{w}_{q+1} - \vec{w}_q)$ is the change in the sum of the estimates from question q to q+1. The remainder of the argument, given in the online appendix, demonstrates that the winner's curse implies that $\vec{e}(\vec{w}_{q+1} - \vec{w}_q)$ is more likely to be positive if $\vec{e}\vec{x}'_{q+1} > 0$, negative if $\vec{e}\vec{x}'_q < 0$, and small in absolute value if $\vec{e}\vec{x}'_{q+1} = 0$. Let us then define "up" questions as those questions in which there are more features on the side of the question predicted to be positive ($\vec{e}\vec{x}'_{q+1} > 0$),[3] "down" questions as those questions in which there are more features on the predicted negative side ($\vec{e}\vec{x}'_q < 0$), and "same" questions as those questions in which there are equal numbers of features on both sides ($\vec{e}\vec{x}'_{q+1} = 0$).

The winner's curse predicts more "cursed" questions for adaptive utility balance than for an algorithm in which utility balance is not a criterion.[4] This will lead to an increase in the average estimated partworths for "up" questions, a decrease for "down" questions, and no change for "same" questions. Because utility balance makes $\hat{a}_{q+1}$ as small as possible, we also expect more "same" questions and more "down" questions.[5]

Smaller partworths are more likely than larger partworths to be updated in the same direction as the sum of the partworths. For example, consider a situation in which exactly three binary partworths vary in a question. With utility balance, $\vec{x}_{q+1}$ will contain either (1) two +1's and one −1 or (2) two −1's and one +1. Because the utility-balance goal is to make $\vec{x}_{q+1}\vec{w}_q$ as small as possible, the smaller partworths are more likely than larger partworths to be in the same direction as $\vec{e}\vec{x}'_{q+1}$. That is, the smaller partworths are more likely to be on the same side as either the two +1's or the two −1's.

[3] Recall that the jth coefficient of x_{q+1} is coded 0, +1, or −1 if the feature is not involved, is present in the right profile only, or is present in the left profile only, respectively.

[4] ACA uses utility balance to select from a set of equally balanced and orthogonal questions. We remove utility balance by selecting randomly from this set.

[5] "Same" questions naturally tend to be more balanced. "Down" questions have more −1's than +1's *and* have predicted positive answers. Such predicted answers are likely to be small in absolute values.

Simulation Evidence for the Winner's Curse

We completed the following four simulation experiments to isolate the winner's curse explanation:

(1) Comparing ACA question selection to a modified version of ACA that does not involve utility balance confirms that the evolution of estimates (q to $q + 1$) has a positive trend for ACA but not for nonutility-balanced ACA. Detailed predictions (described above) are also confirmed.

(2) Restricting ACA question selection to remove the winner's curse ("same" questions) removes the bias.

(3) The overall bias increases as more questions are asked.

(4) Redrawing noise to retain utility balance but remove endogeneity removes the bias.

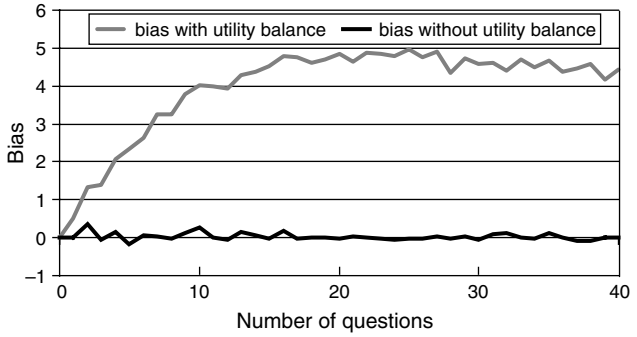
Test 1. The predictions are tested in Table 3 where four features are allowed to vary in each question. (We obtain similar results when we allow two, three, five, six, or seven features to vary. Details are in the online appendix.) As predicted, adaptive utility balance leads to more "same" questions and more "down" questions. For each category ("up," "down," or "same"), more questions are "cursed" when we use adaptive utility balance. "Up" evolution leads to an upward bias (on average partworth estimates increase by 1.25 when utility balance is used but decrease by 0.10 when it is not) and "down" mitigates the upward bias (increase of 1.06 versus increase of 1.25), but the effects do not cancel out. The net effect, for both "up" and "down" questions, is systematically toward

Table 3 Simulations to Demonstrate the Winner's Curse

	Adaptive utility-balanced questions	Questions not utility balanced	Restrict to "same" questions
Percent of "up" questions	17%	59%	—
Percent of "down" questions	14%	4%	—
Percent of "same" questions	69%	37%	100%
Percent of "up" questions that are "cursed"	61%	49%	—
Percent of "down" questions that are "cursed"	17%	14%	—
Percent of "same" questions that are "cursed"	35%	34%	35%
Evolution for "up" questions	1.25	−0.10	—
Evolution for "down" questions	1.06	1.25	—
Evolution for "same" questions	0.06	−0.01	0.00
Evolution for "up" questions x percent "up"	0.22	−0.06	—
Evolution for "down" questions x percent "down"	0.14	0.05	—
Evolution for "same" questions x percent "same"	0.04	−0.01	0.00
Overall bias	4.21*	0.11	−0.25

*Significant at 0.01 level.

Figure 1 Bias as a Function of the Number of Adaptive Utility-Balanced Questions



upward overcorrections in the evolution of the estimates from question q to $q+1$.

Test 2. If the winner's curse is the correct explanation, we can eliminate bias if we restrict ourselves to "same" questions such that $\vec{e}\vec{x}'_{q+1} = 0$. This algorithm eliminates bias (last column of Table 3). This algorithm is of theoretical interest only; it is not designed to be practical. We examine feasible alternatives later in this paper.

Test 3. For the simulations in Table 3, Sawtooth (2002, p. 11) recommends that approximately 20 paired-comparison questions be chosen adaptively from a set of 1,470 possible questions. If all 1,470 questions were asked, we expect the bias to disappear. However, the winner's curse and Equation 1 predict that bias should increase with the number of questions, at least in the beginning. Figure 1 examines this prediction. Consistent with the winner's curse, biases appear to increase for the first 20 or so questions and stabilize with a slight decrease until at least the fortieth question. Figure 1 also cautions that adaptive metric utility balance could give a false impression that true partworths change as more questions are asked. We recommend instead the procedure developed by Liechty et al. (2004) to examine dynamic changes in true partworths.

Test 4. The winner's curse requires both endogeneity (adaptation) and utility balance. Removing endogeneity by redrawing response errors for ACA questions should remove the winner's curse. When we redraw noise for the simulations in Tables 1 and 3, the bias becomes insignificant (-0.48 for the Table 1 questions and 0.20 for the Table 3 questions).

In summary, all four tests are consistent with the winner's curse explanation. Finally, we note that published evidence suggests a statistically significant 6.6% bias for ACA when averaged across domains that include both high and low response error and high and low heterogeneity (Toubia et al. 2003, p. 285).

4. Metric Utility Balance and the Reduction in Efficiency

Perhaps we should accept a modest amount of bias if there are reciprocal benefits. For example, hierarchical Bayesian methods shrink individual estimates toward the population mean to enhance accuracy. In another example, Huber and Zwerina (1996, p. 309, 312) attempt to "improve efficiencies of (choice) designs by balancing the utilities of the alternatives in each choice set," and demonstrate that swapping and relabeling to increase utility balance improves efficiency (defined below) by 33%. Their improved design is significantly more utility balanced than an orthogonal design—11 of the 15 pairs are balanced (versus 0 of the 15 in the orthogonal design). (Their design is, appropriately, not perfectly balanced—an issue we address below.) Thus, if greater utility balance in choice designs increases efficiency, we should explore whether greater utility balance for metric paired-comparison questions increases efficiency. Perhaps increased efficiency could justify the modest bias introduced by the winner's curse.

Efficiency focuses on the standard errors of the estimates. When the estimates, $\hat{\vec{w}}$, of the partworth vectors, $\vec{w}$, are (approximately) normally distributed with variance Σ, the confidence region for the estimates is an ellipsoid defined by $(\vec{w} - \hat{\vec{w}})'\Sigma^{-1}(\vec{w} - \hat{\vec{w}})$ (Greene 1993, p. 190). For most estimation methods, Σ (or its estimate) depends upon the questions that the respondent is asked, and hence, an efficient set of questions minimizes the confidence ellipsoid. This is implemented as minimizing a norm of the matrix, Σ. A-errors are based on the trace of Σ; D-errors are based on the determinant of Σ (Kuhfeld et al. 1994, p. 547).

For metric data, $\Sigma^{-1} = X'X$ (for an appropriately coded X). Perfect *metric* utility balance imposes a linear constraint on the columns of X because $X\vec{w} = \vec{0}$. This constraint induces $X'X$ to be singular. When $X'X$ is singular, the determinant will be zero and D-errors increase without bound. Imperfect metric utility balance leads to smaller $\det(X'X)$ and larger D-errors. A-errors behave similarly. Thus, utility balance is likely to make metric questions inefficient. As an illustration, we computed the average efficiency loss with ACA relative to a fixed orthogonal design for simulations replicating Table 1. The net loss in efficiency was 26.5%.

Unlike choice-based utility balance, greater *metric* utility balance does not seem to lead to increased efficiency.

5. Empirical Issues with Metric Utility-Balanced Questions

Orme (1999, p. 2) hypothesizes that "a difficult choice provides better information for further refining utility

estimates." Huber and Hansen (1986) ascribe "greater respondent interest in these difficult-to-judge pairs." For *choice* data, Haaijer et al. (2000, p. 380) suggest that respondents take more time on choice sets that are balanced in utility, and therefore make less error-prone choices. These hypotheses are consistent with Shugan's (1980) theory that the cost of thinking is inversely proportional to the square of the utility difference. On the other hand, as noted previously, Green et al. (1991, p. 221) hypothesize that respondents tend to use more of the response scale than would be predicted by utility balance.

To investigate whether metric utility balance leads to lower response errors, which might compensate for endogeneity bias and inefficiency, we used HB methods to estimate partworths based on only the paired-comparison questions in the laptop computer bag data (Table 2). HB estimation provides an estimate of the response error in these questions—it was 16% larger for ACA-chosen questions than for fixed, orthogonal paired-comparison questions (significant at the 0.01 level). Thus, in this empirical example, we could find no evidence that metric utility balance led to lower response error than orthogonal questions.[6]

The empirical estimate of response error also gives us the ability to examine the magnitude of endogeneity bias. Simulations suggest that endogeneity bias is approximately 12%–18% and that efficiency losses are approximately 26.5%. As a comparison, empirical data suggest that response error is approximately 21% of total utility. Thus, systematic endogeneity bias and efficiency loss are of the order of magnitude of response error. This is good news. The practical impact of adaptive metric utility balance will only affect those managerial decisions that are highly sensitive to partworth estimates. However, as we argue below, there are good alternatives to metric utility balance. Researchers can obtain the benefits of adaptation and challenging questions without the problems introduced by adaptive metric utility balance.

6. Heterogeneity and Selection Bias in Metric Utility-Balanced Questions

Another hypothesis might be that we can mitigate endogeneity bias with procedures that "borrow information from other respondents" by shrinking individual respondent estimates toward the population mean (Sawtooth Software 2001, p. 1).[7] HB estimation provides a viable alternative that has proven effective in metric conjoint analysis (Lenk et al. 1996). It is easy to verify with a simulation problem similar to that in Table 1 that HB estimates with orthogonal questions produce *average* partworth estimates that are unbiased.

However, were we to attempt a "shrinkage" estimation for utility-balanced questions, our estimate of the population means would suffer from another form of endogeneity bias in question selection (selection bias among respondents). In particular, when questions are adapted to each respondent, the questions are based on the respondent's true partworths as well as the noise in the respondent's answers. As an illustration, reconsider the stylized 2×2 model and suppose that the two partworths are distributed across respondents with probability density function, $f(w_1, w_2)$. For simplicity, suppose there is no measurement noise. We still ask the first utility-balanced question and it is unbiased. In the second question, we encounter Case 1 and observe $\widehat{w}_1 = w_1$ *only* for those respondents for which $w_1 < w_2$. We encounter Case 2 and observe $\widehat{w}_2 = w_2$ *only* for those respondents for which $w_2 < w_1$. Thus, the observations on partworths will be biased toward those respondents with lower partworths.

The selection bias occurs because $E[w_1 \mid w_1 \leq w_2] \leq E[w_1]$ and $E[w_2 \mid w_2 \leq w_1] \leq E[w_2]$. Furthermore, because selection bias depends on the lower tails of the density function, the bias will be greater for higher levels of heterogeneity. We demonstrate this formally in the online appendix for the stylized model—the average observation for the second question is $\overline{w} - \delta/3$, where δ is an index of heterogeneity.

To illustrate the phenomenon with a realistic problem, we repeated the simulations in Table 1, but chose true partworths from a normal distribution and used OLS to obtain aggregate partworths (one regression using data from all respondents). In this case, the *aggregate* partworths were *downwardly* biased by 41% relative to orthogonal questions. Relative selection bias was also significant ($r = 0.93$, $t = 7.3$) and led to smaller normalized mean partworths being biased downward by 30%–50% and larger normalized mean partworths being biased upward by 6%–7% (see the online appendix). Because HB shrinks partworth estimates toward the population mean, it too will be affected by the selection biases in the data. For example, the estimates of the population means, produced by applying HB with ACA questions in a simulation problem similar to that in Table 1, were significantly biased downwards. (Please note that this is a data problem because of adaptive metric utility-balanced question selection, not a problem with HB estimation.)

To illustrate the differential impacts of endogeneity and selection biases, we use four interrelated simulations (see Table 4). We keep ACA questions constant

[6] Aggregate estimates themselves might be subject to selection biases as discussed later in this paper. That phenomenon might cause us to underestimate response errors in ACA. However, at minimum, we can state that there is no evidence in these data to suggest that ACA leads to lower response errors.

[7] Approximately 25% of ACA applications use HB estimation (Sawtooth 2004).

Table 4 Examining the Causes of Endogeneity and Selection Biases

	Endogeneity bias	Selection bias
Adaptive utility-balanced questions	Significant	Significant
Same response errors, heterogeneity redrawn	Significant	Not significant
Response errors redrawn, same heterogeneity	Not significant	Significant
Response error and heterogeneity redrawn	Not significant	Not significant

and redraw either heterogeneity and/or response error. All simulations use OLS to estimate the population means as well as the individual utilities.

It is fortuitous that endogeneity bias *raises* partworth estimates and selection bias *lowers* partworth estimates and both act disproportionally on small partworths. However, it is dangerous to assume that the relative biases will cancel. More importantly, if partworths are differentially heterogeneous (e.g., respondents vary in their preferences for color but not for handles), then aggregate ACA estimates will be biased in unpredictable ways. For example, the average individual respondent-based orthogonal question and ACA question respondent partworths in Table 2 are significantly correlated ($r = 0.75$, $t = 3.2$), but the aggregate estimates are not significantly correlated ($r = 0.39$, $t = 1.2$) (details in the online appendix). Indeed, the aggregate ACA estimates are 53% lower than the orthogonal question-based estimates. Thus, alas, shrinkage estimates do not overcome the biases in question selection introduced by adaptive metric utility balance.

7. Alternatives to Utility Balance for Adapting Metric Questions

If a researcher wishes to retain metric questions, then an alternative criterion exists to metric utility balance. Polyhedral methods select questions to reduce the feasible set of partworths rapidly by focusing questions relative to the "axis" about which there is the most uncertainty. This criterion does not appear to be subject to a winner's curse. Furthermore, this criterion imposes a constraint that the rows of X be orthogonal (Toubia et al. 2003, Equation A9). After p questions, X will be square, nonsingular, and orthogonal ($XX' \propto I$ implies $X'X \propto I$).[8] Subject to scaling, this orthogonality relationship minimizes D-error (and A-error). Thus, while the adaptation inherent in metric polyhedral methods leads to endogenous question design, the lack of an explicit winner's curse and the orthogonality constraint appear to avoid the biases in Tables 1–3. Toubia et al. (2003) report at most a 1% bias for metric polyhedral question selection, significantly less than observed for ACA question selection. Polyhedral question selection performs better than ACA in simulation (mean absolute error of true versus predicted partworths) and as well or better than ACA in the two empirical tests to date. Toubia et al. (2003, Table 7) report significantly better performance when almost all questions were chosen by polyhedral methods. Orme and King (2002) compare ACA to a hybrid in which one-third of the questions are polyhedral and two-thirds of the questions are ACA. They report no significant differences. Thus, polyhedral methods do at least as well as ACA question selection and represent a viable alternative to metric utility balance. Other researchers are working on another alternative question-selection algorithm that is based on support vector machines (Evgeniou et al. 2004).

8. Alternatives to Metric Questions—Utility Balance for Choice-Based Questions

Arora and Huber (2001), Huber and Zwerina (1996), and Kanninen (2002) use utility balance as *one* criterion for choice-based questions. They improve D_p-efficiency relative to orthogonal designs by using designs that are significantly more utility balanced—e.g., 73% utility balanced for customized versus 0% utility balanced for orthogonal in Huber and Zwerina (1996, Table 3).[9] Choice-based utility balance can improve efficiency because the covariance matrix of the choice-based estimates is a function of the choice probabilities. Thus, if providing challenging questions to respondents is our goal, we can use choice questions rather than metric questions. In the online appendix, we show formally that, as response errors decrease, optimal questions become more utility balanced. Proposition 2 is consistent with prior simulations. For example, in Huber and Zwerina (1996, Table 2), more swaps are accepted (greater utility balance) when response accuracy increases. Arora and Huber (2001, Table 2) also report greater efficiency gains for more accurate partworths.

Proposition 2. *When optimizing A-efficiency for binary choice, greater response accuracy implies greater utility balance.*

Utility balance could also be helpful when designing questionnaires to elicit respondents' reservation prices for product bundles (see Jedidi et al. 2003) and when designing choice experiments to augment panel data (see Swait and Andrews 2003).

[8] Orthogonal rows assume that $XX' = \alpha I$, where α is a proportionality constant. Because X is nonsingular, X and X' are invertible, thus $X'XX'X'^{-1} = \alpha X'X'^{-1} \Rightarrow X'X = \alpha I$.

[9] Endogeneity bias is not a problem in these algorithms; they do not adapt questions for individual respondents.

9. Adaptive Questions for Choice-Based Data

Finally, we examine whether we introduce bias by using utility balance to adapt choice-based questions based on individual respondent data. We examined relative bias and efficiency for one such algorithm based on polyhedral methods (Toubia et al. 2004). (Published simulations have already shown that, in most domains, adaptive polyhedral choice-based questions lead to more accurate partworths than either orthogonal or aggregately customized questions.) To test relative bias, we sort the true partworths, divide them into M ordered subsets, and compute the average error (predicted minus true) within each subset. This allows examining bias as a function of the relative size of the partworths. We did this for (1) adaptive polyhedral questions and (2) orthogonal questions. If there were relative bias, then (1) would be significantly different from (2). They were not significantly different. (Graphs are provided in the online appendix to this paper.) We also computed the average efficiency for (1) and (2). The D_p-efficiency for polyhedral choice questions was 3% higher than for fixed questions, reflecting the fact that efficiency is a function of the true parameters and suggesting that adaptive polyhedral questions do not lead to any loss in efficiency relative to orthogonal questions.

10. Summary

We examine the endogeneity bias, lowered efficiency, response errors, and selection biases that result from adaptive metric utility balance—the question selection algorithm at the heart of ACA. We have shown that the biases and inefficiencies are real and in the direction predicted. We provide stylized models and more general explanations with which to understand and isolate the cause of these phenomena. Furthermore, empirically, we find no evidence that metric utility-balanced questions reduce response error. Contrary to common wisdom, orthogonality (efficiency) in metric questions appears to be a more important goal than utility balance.[10]

Fortunately, the adverse effects of adaptive metric utility balance can be avoided easily. For those researchers seeking to retain metric questions, polyhedral methods provide an alternative that avoids both endogeneity bias and lowered efficiency. For those researchers seeking to ask difficult, challenging questions to encourage respondents to think harder, utility balance appears to help choice-based questions. Finally, adaptive utility-balanced choice questions do not appear to be biased.

[10] In a related paper (available from the authors), we examine another form of efficiency designed to enhance managerial decisions by focusing on managerially relevant combinations of partworths, $M\vec{w}$. We examine the properties of M-efficiency, which minimizes a norm of $M(X'X)^{-1}M'$.

Acknowledgments

This research was supported by the MIT Sloan School of Management and MIT's Center for Innovation in Product Development. An appendix to this paper contains additional analyses, tables, figures, and proofs to the propositions and may be downloaded from http://mktsci.pubs.informs.org. Demonstrations of many of the preference measurement questions discussed in this paper can be viewed at http://mitsloan.mit.edu/vc. Special thanks to Eric Bradlow, Ely Dahan, Theodoros Evgeniou, Jim Figura, Sangman Han, Oded Koenigsberg, Oded Netzer, Duncan Simester, and the *Marketing Science* reviewers, area editor, and editor-in-chief for insightful comments and examples. The authors are listed in alphabetical order; contributions were equal and synergistic.

References

Allenby, Greg M., Neeraj Arora. 1995. Incorporating prior knowledge into the analysis of conjoint studies. *J. Marketing Res.* **32**(2) 152–163.

Arora, Neeraj, Joel Huber. 2001. Improving parameter estimates and model prediction by aggregate customization in choice experiments. *J. Consumer Res.* **28**(September) 273–283.

Capen, E. C., R. V. Clapp, W. M. Campbell. 1971. Competitive bidding in high-risk situations. *J. Petroleum Tech.* **23**(June) 641–653.

Carroll, J. Douglas, Paul E. Green. 1995. Psychometric methods in marketing research: Part I, Conjoint analysis. *J. Marketing Res.* **32**(November) 385–391.

Choi, S. Chan, Wayne S. DeSarbo. 1994. A conjoint-based product designing procedure incorporating price competition. *J. Product Innovation Management* **11** 451–459.

Evgeniou, Theodoros, Constantinos Boussios, Giorgos Zacharia. 2004. Generalized robust conjoint estimation. *Marketing Sci.* **24**(3) 415–429.

Green, Paul E., Abba Krieger. 1995. Attribute importance weights modification in assessing a brand's competitive potential. *Marketing Sci.* **14**(3, Part 1 of 2) 253–270.

Green, Paul E., Abba Krieger, Manoj K. Agarwal. 1991. Adaptive conjoint analysis: Some caveats and suggestions. *J. Marketing Res.* **28**(2) 215–222.

Greene, William H. 1993. *Econometric Analysis*, 2nd ed. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Haaijer, Rinus, Wagner Kamakura, Michel Wedel. 2000. Response latencies in the analysis of conjoint choice experiments. *J. Marketing Res.* **37**(August) 376–382.

Hauser, John R., Steven M. Shugan. 1980. Intensity measures of consumer preference. *Oper. Res.* **28**(2, March–April) 278–320.

Huber, Joel, David Hansen. 1986. Testing the impact of dimensional complexity and affective differences of paired concepts in adaptive conjoint analysis. Melanie Wallendorf, Paul Anderson, eds. *Advances in Consumer Research*, Vol. 14. Associations of Consumer Research, Provo, UT, 159–163.

Huber, Joel, Klaus Zwerina. 1996. The importance of utility balance in efficient choice designs. *J. Marketing Res.* **33**(August) 307–317.

Huber, Joel, Dick R. Wittink, John A. Fiedler, Richard Miller. 1993. The effectiveness of alternative preference elicitation procedures in predicting choice. *J. Marketing Res.* **30**(1) 105–114.

Jedidi, Kamel, Sharan Jagpal, Puneet Manchanda. 2003. Measuring heterogeneous reservation prices for product bundles. *Marketing Sci.* **22**(1, Winter) 107–130.

Johnson, Richard. 1991. Comment on adaptive conjoint analysis: Some caveats and suggestions. *J. Marketing Res.* **28**(May) 223–225.

Johnson, Richard, Joel Huber, Lynd Bacon. 2003. Adaptive choice-based conjoint. Sawtooth Software, Sequim, WA.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, T. C. Lee. 1985. *The Theory and Practice of Econometrics*. John Wiley and Sons, New York.

Kagel, John H., Dan Levin. 1986. The winner's curse and public information in common value auctions. *Amer. Econom. Rev.* **76**(5, December) 894–920.

Kanninen, Barbara J. 2002. Optimal design for multinomial choice experiments. *J. Marketing Res.* **39**(May) 214–227.

Kuhfeld, Warren F., Randall D. Tobias, Mark Garratt. 1994. Efficient experimental design with marketing research applications. *J. Marketing Res.* **31**(4, November) 545–557.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, Martin R. Young. 1996. Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Sci.* **15**(2) 173–191.

Liechty, John, Duncan Fong, Wayne S. DeSarbo. 2005. The evolution of consumer utility functions in conjoint analysis. *Marketing Sci.* **24**(2) 285–293.

Marshall, Pablo, Eric T. Bradlow. 2002. A unified approach to conjoint analysis models. *J. Amer. Statist. Assoc.* **97**(459, September) 674–682.

Orme, Bryan. 1999. ACA, CBC, or both: Effective strategies for conjoint research. Working paper, Sawtooth Software, Sequim, WA.

Orme, Bryan, Joel Huber. 2000. Improving the value of conjoint simulations. *Marketing Res.* **12**(4, Winter) 12–20.

Orme, Bryan, W. Christopher King. 2002. Improving ACA algorithms: Challenging a twenty-year-old approach. *Advance Res. Tech. Conf.* American Marketing Association, Vail, Co.

Sandor, Zsolt, Michel Wedel. 2001. Designing conjoint choice experiments using managers' prior beliefs. *J. Marketing Res.* **38**(4, November) 430–444.

Sandor, Zsolt, Michel Wedel. 2002. Profile construction in experimental choice designs for mixed logit models. *Marketing Sci.* **21**(4, Fall) 455–475.

Sándor, Zsolt, Michel Wedel. 2003. Differentiated Bayesian conjoint choice designs, April 29, 2003. ERIM Report Series Reference No. ERS-2003-016-MKT. http://ssrn.com/abstract=41161.

Sawtooth Software. 2001. The ACA/hierarchical Bayes technical paper. Sawtooth Software, Inc., Sequim, WA.

Sawtooth Software. 2002. ACA 5.0 technical paper. Sawtooth Software Technical Paper Series, Sawtooth Software, Inc., Sequim, WA.

Sawtooth Software. 2004. Update on relative conjoint analysis usage. *Sawtooth Solutions* (Summer). http://www.sawtoothsoftware.com/productforms/ssolutions/ss20.shtml#ss20usage.

Shugan, Steven M. 1980. The cost of thinking. *J. Consumer Res.* **7**(2, September) 99–111.

Swait, Joffre, Rick L. Andrews. 2003. Enriching scanner panel models with choice experiments. *Marketing Sci.* **22**(4) 442–460.

Thaler, Richard H. 1992. *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. Princeton University Press, Princeton, NJ.

Toubia, Olivier, John R. Hauser, Duncan I. Simester. 2004. Polyhedral methods for adaptive choice-based conjoint analysis. *J. Marketing Res.* **41**(1) 116–131.

Toubia, Olivier, Duncan Simester, John R. Hauser, Ely Dahan. 2003. Fast polyhedral adaptive conjoint estimation. *Marketing Sci.* **22**(3) 273–303.