Sawtooth Software

RESEARCH PAPER SERIES

Common Scale Hybrid Discrete Choice Analysis:

Fusing Best-Worst Case 2 and 3

Bryan Orme
Sawtooth Software, Inc.

1457 E 840 N Orem, Utah
+1 801 477 4700
www.sawtoothsoftware.com

Common Scale Hybrid Discrete Choice Analysis:

Fusing Best-Worst Case 2 and 3

Bryan Orme, Sawtooth Software
February, 2013 (updated Nov. 4, 2013[1])

Abstract:

We summarize previous research regarding the use of Best-Worst Case 2 (so-called *Best-Worst Conjoint*) compared to CBC. One of the biggest advantages of Best-Worst Case 2 is that the utility scores are placed on a common scale with a common origin, allowing one to directly compare every level to every other level, even across attributes. Using synthetic data, we compare the statistical efficiency of Best-Worst Case 2 to CBC. Next, we introduce a new hybrid method that fuses prior rank order information for attribute levels (via soft constraints), Best-Worst Case 2, and Best-Worst Case 3 (also known as *Best-Worst CBC*). Finally, we report on an empirical test of our proposed hybrid approach versus CBC. We find that Best-Worst Case 2 utilities are quite similar to but statistically different from CBC utilities. CBC utilities predict CBC holdout tasks better than Best-Worst Case 2 or fusions involving Best-Worst Case 2 and Best-Worst Case 3 (Best-Worst CBC). Best-Worst Case 2 is not a substitute for CBC, but the advantage of the common utility scale could be quite useful for such research contexts as messaging, employee research, and healthcare outcomes. Best-Worst Case 2 involves a different preference elicitation task from CBC wherein respondents state which levels are most and least important, or most and least preferred. CBC, in contrast, derives via statistical analysis which levels are driving preference by analyzing respondent choices of whole product profiles. Researchers employing the proposed hybrid discrete choice analysis approach could obtain nearly identical results as CBC (for market simulations) by dropping the Best-Worst Case 2 section during the analysis, while achieving the benefits of the common scale (for strategic interpretation and market segmentation) by including the Best-Worst Case 2 information within the fusion.

Introduction:

Stated choice experiments (discrete choice or CBC) have become popular for studying buyers' preferences for products and services. The Sawtooth Software user community has embraced CBC (Louviere and Woodworth 1983, Sawtooth Software 1993), making it the most-often used conjoint analysis approach since about 2000 (displacing ACA which previously was the most popular conjoint technique).

Over the last 30+ years, researchers have employed many variations on stated choice experiments, including:

- Discrete choice (the respondent picks the best alternative within a set of alternatives)
- Allocation-based choice (the respondent allocates a fixed number of points across alternatives within the set, or volumetric CBC where the allocated points do not have to sum to a fixed value)

[1] Updated November 4, 2013 to include reference to the investigation of Flynn *et al*. 2013.

- Best-Worst CBC, also known as Best Worst Case 3 (the respondent selects both the best and worst product alternatives within a set of alternatives)
- Dual-response None CBC, where respondents are forced to pick a product and then asked in a follow-up question if they really would buy the product they just selected (Uldry *et al.*, 2002)
- Adaptive CBC (Sawtooth Software 2008)
- Menu-Based Choice (Sawtooth Software 2012)

A related choice method, Maximum Difference Scaling (MaxDiff, also known as Best-Worst Scaling), is a technique invented by Jordan Louviere in 1987 while on the faculty at the University of Alberta (Louviere, Personal Correspondence 2005). The first working papers and publications occurred in the early 1990s (Louviere 1991, Finn and Louviere 1992, Louviere 1993, Louviere, Swait, and Anderson 1995). Louviere and colleagues used best-worst for obtaining importance measurements across an array of items (known as Best-Worst Case 1). Sawtooth Software's MaxDiff software (Sawtooth Software 2004) is based on Louviere's work. The vast majority of MaxDiff studies today are of the Case 1 type.

Louviere and colleagues also proposed that MaxDiff could be used similar to conjoint analysis if the array of items was like the attributes and levels used to comprise a conjoint profile (multiple attributes each with mutually exclusive levels) (Louviere, Swait, and Anderson 1995, Louviere 1994, Marley, Flynn, and Louviere 2008). They called this approach *Best-Worst Conjoint*, and later *Best-Worst Case 2*. With Best-Worst Case 2, an array of conjoint profiles is constructed as is done when designing conjoint experiments with Sawtooth Software's CVA system for card-sort conjoint (e.g. typically a dozen to thirty product profiles, following a fractional factorial, level-balanced, near-orthogonal plan). With Best-Worst Case 2, the respondent evaluates each profile, but rather than express preference for the conjoined set of attribute levels comprising the profile, the respondent states which one attribute level within each profile makes the respondent most want to purchase the product and which one attribute level makes the respondent least want to purchase the product.

Many researchers, including Sawtooth Software's founder Rich Johnson, recoiled at the notion that Best-Worst Case 2 could be used as if it were conjoint analysis. After all, respondents never express preference for the product profiles as conjoined wholes (as a rating, ranking across profiles, or choice among a set of profiles). Therefore, there was no compositional rule, meaning that the dependent variable is not predicted by a linear expression involving multiple summed attributes. Thus, Best-Worst Case 2 was not formally a conjoint method. Despite this controversy, Louviere advocated (Louviere, Swait, and Anderson 1995) that the item scores estimated using Best-Worst Case 2 could be summed to construct preferences for multi-attribute product profiles and used within choice simulators similar to CBC or conjoint part-worths[2]. Another aspect that made Best-Worst Case 2 less compelling than conjoint techniques is that it asked respondents to directly state which specific attribute levels were driving their decisions, rather than observing overall choices of product concepts and deriving the preferences using methods such as regression analysis, MNL, or HB. It seems much more likely that socially desirable responses could bias respondents' choices within Best-Worst Case 2 than for conjoint analysis. Despite these drawbacks, what made Best-Worst Case 2 unique and potentially valuable was

[2] A related claim is made by Srinivasan and Park in their paper "The Surprising Robustness of Self-Explicated Models" (Srinivasan and Park 1997) where they demonstrate that self-explicated scores (also lacking a compositional rule) may be used within conjoint simulators and perform essentially as well in predicting holdout choices (even better for their application) as conjoint utilities.

that *it led to scores on a common scale*, permitting direct comparison between all attribute levels within the study, not just direct comparisons among levels within the same attribute (as is the case for other conjoint methods).

Besides Louviere and co-authors, other researchers have also reported that Best-Worst Case 2 can predict CBC holdouts as accurately as CBC, though the part-worth parameters are not equivalent to CBC (Chrzan and Skrapits, 1996, Chrzan and Loscheider 2013). We present new research within this article also finding that the Best-Worst Case 2 parameters are not equivalent to CBC, but *not* confirming that they can predict CBC holdouts as well as CBC (though they do perform reasonably well).

Some Previous Research Comparing Best-Worst Case 2 and CBC

In a 1995 working paper by Louviere, Swait, and Anderson, the authors compared Best-Worst Case 2 to CBC for a study involving choice of ski resorts, with a 13 attribute design involving 48 total levels (2^2 x 4^{11}). Sample sizes were 282 respondents for CBC, 195 for Best-Worst Case 2, and 152 for a revealed choice format. Respondents received 16 profiles within the Best-Worst Case 2 cell. Respondents completed 10 CBC tasks (2 concepts each) within the CBC cell. The authors fit aggregate MNL models and found no significant differences between the CBC and Best-Worst Case 2 parameters (after accounting for differences in scale, via the Swait-Louviere test). The scale factor was larger for Best-Worst Case 2 than CBC, implying lower response error for Best-Worst Case 2 (a finding replicated by us and other authors). Of the two methods, the authors state (p 19): "…differences between methods were mainly due to reliability differences (i.e. error variability) in measuring the common underlying preferences, not differences in cognitive processes… Our results also imply that one could combine these different data sources to obtain more precise estimates and more discriminating tests of behavioral hypotheses." They also state: "BW [Best-Worst Case 2] and CB conjoint [CBC] can be used separately or as complementary techniques". They further indicate that the data could be fused with buy/no buy (dual-response None tasks) to estimate a None parameter: "Non-choice probabilities also can be modeled by jointly estimating models from BW and choice/non-choice (yes, no) data and accounting for scale differences".

In 1996, Chrzan and Skrapits compared Best-Worst Case 2 to CBC within the context of a study involving a technology product with 9 attributes (16 x 6 x 5^5 x 3^2). Respondents completed both Best-Worst Case 2 and CBC questionnaires. For the Best-Worst Case 2 questionnaire, 4 profiles were shown to each respondent who was asked to select the best and worst levels within each profile. The CBC questionnaire included 6 choice tasks. Multiple blocks (questionnaire versions) were distributed across respondents (for both Best-Worst Case 2 and CBC cells), to permit robust estimation of part-worth parameters[3]. Aggregate MNL was used for analysis (this was in the days before the general availability of HB analysis). The aggregate MNL parameters between Best-Worst Case 2 and CBC were correlated at 0.90[4]. Holdout prediction (shares of choice for CBC-looking tasks) was slightly better for Best-Worst Case 2, though the authors admit that the design of the holdout choices was unusual. The parameters were not equivalent per the Swait-Louviere test (p<.01). Lack of equivalence of the parameters meant

[3] With this relatively large design, 44 part-worth parameters were estimated.

[4] The Best-Worst Case 2 data were zero-centered within factor (by subtracting off the mean part-worth utility for the levels within each factor from each level) prior to running the correlation analysis with the CBC part-worth utilities.

that the scale for the two methods could not be accurately compared in a formal sense. In terms of absolute magnitude, the Best-Worst Case 2 effects were about 3x the size of the CBC effects, suggesting much lower response error for the Best-Worst Case 2 data.

In 2013, Chrzan and Loscheider reported results of a split-sample experiment comparing Best-Worst Case 2 to CBC for a study involving refrigerators on just 4 attributes with 3 levels each (3^4) (Chrzan and Loscheider 2013). Respondents received 9 Best-Worst Case 2 profiles or 9 CBC tasks each with 3 concepts (minimal overlap). They included a cell of holdout respondents in the experiment who each received 9 holdout CBC tasks (4 concepts each, minimal overlap). They estimated parameters for Best-Worst Case 2 and CBC cells using both aggregate MNL and HB. The aggregate MNL parameters were correlated 0.90, however the Swait-Louviere test rejected the null hypothesis (p<.01) that the two sets of parameters were equivalent after adjusting for scale. The magnitude of the Best-Worst Case 2 utilities seemed about double that of the CBC utilities. Holdout predictions of CBC tasks both for aggregate MNL and HB (correlations with aggregate shares, tuned individually for scale to minimize MAE) slightly favored Best-Worst Case 2, even though methods bias favored CBC for predicting CBC holdouts. The MAE for the two approaches was 0.028 for Best-Worst Case 2 and 0.034 for CBC. The correlation between predictions and actual holdout choice shares was 0.945 for Best-Worst Case 2 and .941 for CBC. The authors concluded: “BW-C2 seems like a viable methodology for measuring part-worth utilities for conjoint analysis models with generic attributes.”

Also in 2013, Flynn, Peters and Coast reported on a study fielded (regarding quality of life issues) from 2005-2006 that synthesized both Best-Worst Case 2 with CBC (binary choice of a conjoint profile versus a status quo constant alternative) within the same questionnaire (Flynn *et al*. 2013). They found only weak evidence to reject the null hypothesis that the Best-Worst Case 2 utilities were different from CBC utilities after adjusting for scale. However, they found a rather large difference in the error rate (scale) for the two exercises, suggesting Best-Worst Case 2 data had about seven times less response error than their binary choice CBC exercise. Upon further analysis, Flynn *et al*. found evidence that many respondents did not seem to understand their CBC exercise. They also suggested that a CBC exercise with more concepts (profiles) per task than they used would have higher d-efficiency and may prove better for comparing the differences between the two types of choice exercises.

Synthetic Data to Compare Best-Worst Case 2 and Best-Worst Case 3 (BW-CBC)

As we reported in the previous section, previous research with real respondents has shown that Best-Worst Case 2 and CBC can lead to similar results, both in terms of utility scores and in terms of predicted shares of preference for multi-attribute products. However, we are not aware of previous research that has compared the two methodologies in terms of design efficiency: how many choice tasks are needed to stabilize parameter estimates at the individual level for multi-attribute (conjoint-style) designs. A few authors have investigated how many CBC tasks are needed with real respondents for robust results (Orme and Johnson 1996, Hoogerbrugge 2007, Tang 2010, Kurz and Binner 2012). The authors that employed HB estimation (Hoogerbrugge, Tang, Kurz and Binner) have concluded that results for real respondents don’t benefit much from completing more than about 10 or 12 choice tasks[5].

[5] A large proportion of real respondents use non-compensatory decision rules, rather than additive compensatory, and it can often take only a handful of tasks to identify the key levels such respondents are using to make choices.

To test the relative efficiency of CBC vs. Best-Worst Case 2, we selected a typically sized design as used in practice, with six factors having different numbers of levels each (3^2x4^2x5^2). We constructed two synthetic data sets of 500 respondents, one for CBC and one for Best-Worst Case 2, with part-worth utilities normally distributed (variance=2) around a population mean vector of [-1, 0, 1; -2, 0, 2; -1, -1, 1, 1; -4, -1, 1, 4; -2, -1, 0, 1, 2; -0.50, -0.25, 0, 0.25, 0.50]. Based on our experience, this approximately reflects the magnitude and variance of CBC part-worths as found with real respondents in real CBC data sets. For both experimental designs, we generated 20 choice tasks using one of Sawtooth Software's randomized CBC design strategies (Complete Enumeration) which controls for level balance, orthogonality, and minimal overlap. The CBC questionnaire was constructed with 4 concepts per task, as would be typically done in practice. Computer-generated respondents answered each questionnaire according to true individual-level utilities plus random error (IID Gumbel errors added to the concept utility sums for CBC choices; and IID Gumbel errors added to the true level utilities for the Best-Worst Case 2 choices of best and worst levels within profiles)[6]. Because previous research (plus the new research we present later in this document) has found that Best-Worst Case 2 involves half to 2/3 lower response error than CBC choices, we also investigated the recovery of utility parameters for Best-Worst Case 2 assuming respondents answered with half the response error (Gumbel error * 0.5).

We used Sawtooth Software's CBC/HB system with 5K initial iterations[7] followed by 5K used iterations, (prior variance=2) to estimate scores for both the 500 CBC respondents and the 500 Best-Worst Case 2 respondents[8]. Proper prior covariance matrices were specified per the recommendations of Lenk (Lenk and Orme, 2009). The estimated utilities (point estimates)[9] were compared to the true utilities for each respondent. The root mean square error (RMSE) was computed for each individual and then averaged across individuals. The RMSE under different numbers of tasks and questionnaire types we tested are shown in Exhibit 1 and then smoothed using regression functions in Exhibit 2 below.

[6] Right-skewed Gumbel error was applied to the true level utilities for making the (synthetic) respondent's best choice (level with the highest true utility plus error was selected). Independently drawn left-skewed Gumbel error was applied to the true level utilities for making the respondent's worst choice. The occasional pair of choices of the same level (same level picked as both best and worst) was permitted.

[7] In the interest of saving time, we used fewer iterations than default settings in CBC/HB software. These artificial datasets are robust and quite well-behaved, so convergence is obtained quickly. Further iterations would not have materially changed the findings.

[8] The "best-worst" (2 task) design specification was used rather than the "Maximum Difference" (1 task) specification (expressed in each row as a difference between two levels). Each respondent was considered to have made independent choices (best choices and worst choices) within each profile.

[9] The utilities for Best-Worst Case 2 were post estimation zero-centered within each attribute. Because differences in an intercept for an attribute would factor out within logit-based market simulation results, this is a fair comparison when considering how well the two methodologies could predict market scenario shares of preference for multi-attribute products. Additionally, true utilities, Best-Worst Case 2 utilities, and CBC utilities were all standardized at the individual level to have the same variance, to remove the artifact of arbitrary scale differences, prior to computing RMSE.

Exhibit 1

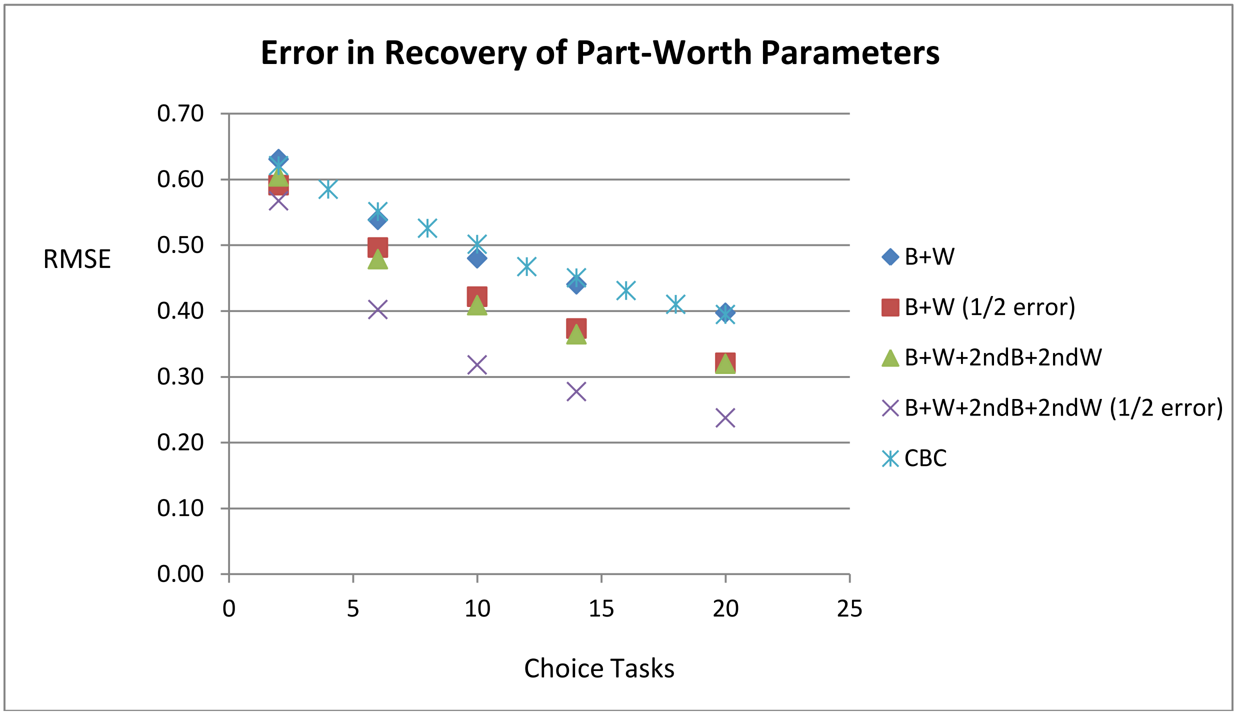
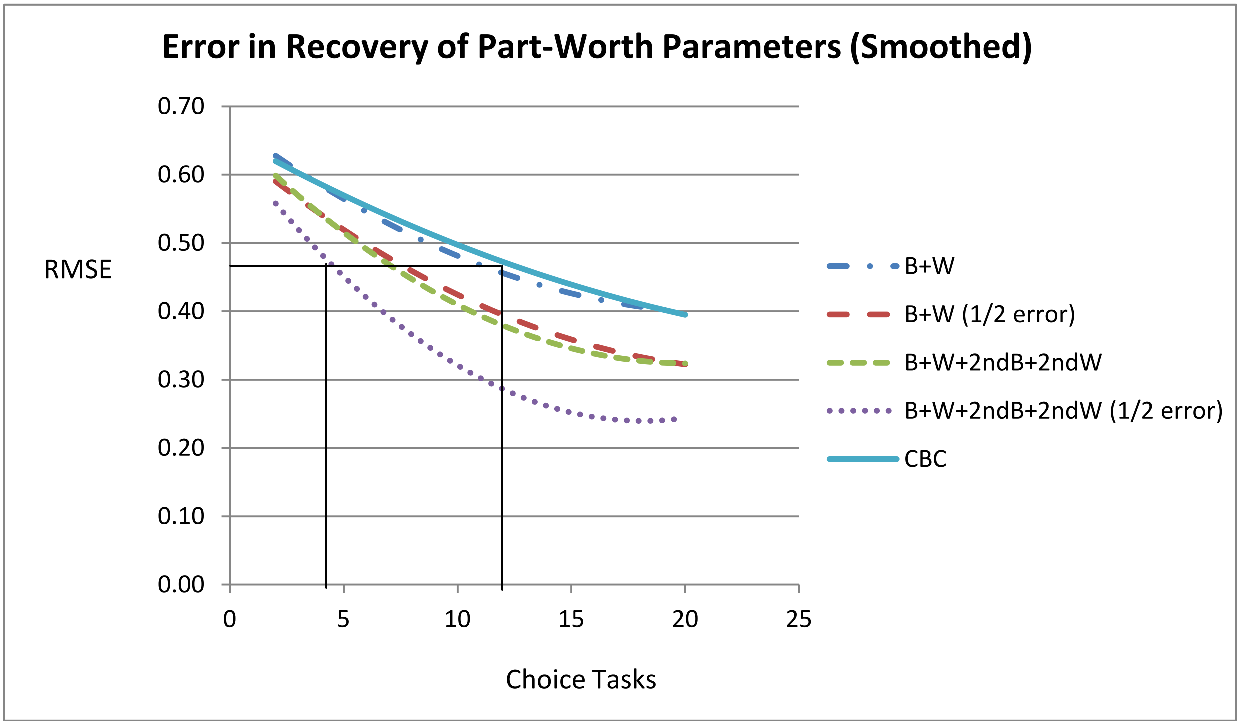


Exhibit 2



When interpreting the results, it's important to keep in mind that a choice task for a Best-Worst Case 2 questionnaire involves showing respondents a single profile (but requiring multiple clicks of best and worst levels); whereas a choice task for CBC involves showing respondents four profiles (but requiring a single click).

As expected, increasing the number of tasks per respondent reduces the error in recovering individual-level part-worth utilities. RMSE to recover true utilities is lower for Best-Worst Case 2 questionnaires, given the same number of choice tasks. An interesting aspect to note is that Best-Worst Case 2 struggles with individual-level estimation when only using 2 choice tasks, as there is a lack of connectivity across the items within each respondent and only 12 of the 24 attribute levels are shown to each respondent.

Generally, about twelve tasks are considered adequate to estimate individual-level utilities for CBC questionnaires for real respondents, given the number of attributes and levels we used in this design. (We'll therefore use that as a benchmark, with RMSE of 0.47 as shown in Exhibit 2.) For our synthetic data set, the same error rate can be achieved using 4.4 Best-Worst Case 2 tasks (where best, worst, 2nd best, and 2nd worst levels are elicited per task and response error is assumed to be half the response error for CBC questionnaires). Very conservatively speaking, the same error rate as 12 CBC tasks could be achieved using 10.7 Best-Worst Case 2 tasks (where 2 clicks are elicited per task and response error is assumed to be equal to the response error for CBC questionnaires). This degree of response error for Best-Worst Case 2 would be much higher than all evidence to date suggests would be the case.

A key take-away is that fewer Best-Worst Case 2 tasks are required relative to CBC to obtain equal precision. For this particular dataset (assuming respondents answer Best-Worst Case 2 questionnaires with half the error and provide 4 clicks per task), slightly fewer than half as many Best-Worst Case 2 tasks are needed relative to CBC tasks to obtain equal precision. Out of necessity (to enable direct comparisons of statistical efficiency) this analysis assumes that the underlying preference scores revealed via CBC or Best-Worst Case 2 are identical, which previous research (plus our new research that we present below) indicates is not the case.

Previous Research Comparing Best-Worst Case 3 and CBC

There has been increasing interest in asking respondents to select both best and worst *concepts* within CBC tasks (Best-Worst Case 3). A few recent synthetic and real (split-sample with human respondents) studies have found modest advantages for Best-Worst Case 3 over CBC (Chrzan *et al.* 2010, Marshal *et al.* 2010, and Wirth 2010). However, another study showed no advantage for worst choices in addition to best choices (Lattery and Orme 2012). We have examined first-choice only versus Best-Worst CBC choices for the same datasets and found that the inclusion of worst choices does not bias the utilities (though it often damps the scale) for these datasets (Orme 2010).

Even though asking worst choices within CBC questions seems to add value in many cases, it isn't clear from previous research whether equally good or better results would result by asking the respondent to make a few more first-choice responses rather than eliciting both best and worst choices within each choice task.

A Hybrid Choice Model

Now that we have compared Best-Worst Case 2, Best-Worst Case 3, and CBC questionnaire approaches, we introduce and test a hybrid conjoint/choice approach that combines Best-Worst Cases 2 and 3 within a single questionnaire. We think our hybrid conjoint method would be useful for designs with somewhere up to 8 or so attributes and about 2 to 7 levels for each attribute. This hybrid approach incorporates three preference elicitation and discrete choice methods that have proven useful over the last decades:

> **Section 1:** Priors Section: Self-explicated ratings of levels within attributes (only for attributes such as brand, style, and color that do not have *a priori* known level order)
>
> **Section 2:** Best-Worst Choices of Levels within Alternatives (Best-Worst Case 2)
>
> **Section 3:** Best-Worst Choices of Alternatives within Sets (Best-Worst Case 3), also known as *Best-Worst CBC*, including optional Dual-Response None questions.

Previous researchers have recommended the fusion of different preference elicitation methods such as stated preference and revealed preference (Hensher, Louviere, and Swait 1999, Louviere et al 1999, Swait, Louviere, and Williams 1994, Ben-Akiva and Morikawa 1990, Hensher and Bradley 1992), self-explicated and conjoint tradeoffs (Sawtooth Software's ACA as well as Paul Green's hybrid conjoint), or BYO (build your own) and CBC tasks (Sawtooth Software's ACBC 2008, Otter 2007). Even though the different preference elicitation contexts involve different degrees of response error, methods have been proposed within logit-based estimation for accounting for differences in scale when fusing the results (Otter 2007). Though they usually lead to slightly different part-worths (after adjusting for scale), the different sections lead to utilities that are usually very highly correlated (often 0.9 or higher), and some researchers have expressed that there is strength in combining data from diverse preference elicitation tasks in studying consumer behavior and for making predictions of market choices. In terms of treating our respondents well, a conjoint interview that offers variety can be more enjoyable than a monotonous, single choice context that is repeated many times—especially if the overall length of the survey is not much longer.

Our hybrid method combines all three preference elicitation sections into a single logit-based maximum likelihood estimation that may be performed using aggregate logit, latent class, or HB[10]. Even though the priors (Section 1) are asked on a ratings-scale, such as a 4-point scale, the data are used only in an ordinal sense and are encoded for utility estimation as choice tasks: pairwise choices between levels of the same attribute, with choices inferred from the ratings data[11]. The final raw utilities may be multiplied by a constant so as to best fit the Best-Worst Case 3 (Best-Worst CBC) choices (Section 3). The Priors (Section 1) and Best-Worst Case 2 (Section 2) choices contain lower response error and thus have too large of a scale to predict the likelihood of choices for standard CBC scenarios optimally.

Why Propose Yet Another Conjoint Approach?

Despite the benefits and wide adoption of conjoint analysis methods, some researchers and clients find it challenging (and limiting) to interpret part-worths, since only within-attribute comparisons are possible. Therefore, a new conjoint method that places all attributes on a common scale and allows direct comparisons across all attributes offers something unique and might make conjoint accessible

[10] The relative weight of the sections may be manipulated by asking respondents to complete more or fewer tasks for sections 2 and 3, by manipulating the magnitude of the X values in the utility estimation matrix, or by replicating certain tasks within the utility estimation matrix.

[11] This approach was suggested by Kevin Lattery (Lattery 2009).

(and appealing) to even more people. One might argue that cross-attribute comparisons for most conjoint analysis studies would seem unnecessary, but we can readily come up with circumstances such as messaging, employee research (job elements), and healthcare[12] applications where cross-attribute comparisons could be especially valuable.

At the 2012 Sawtooth Software conference, Saigal and Dahan showed seven attributes each with two or three levels (see Exhibit 3) that they developed based on intensive personal interviews with patients diagnosed with prostate cancer. Traditional conjoint analysis would not allow the direct comparison of preferences between levels of different attributes. However, it seems reasonable that users of the data might want to know which level was least desirable in terms of treatment outcomes for prostate cancer: "Cutting: surgery with some risks and hospital time" (Attribute 3 level 2) vs. "Urinary: short term issues" (Attribute 5 level 2).

Exhibit 3:
Attributes for Prostate Cancer Treatment Conjoint Analysis

1) OTHERS SUPPORT:
- Doctor and family support this treatment
- Doctor and family do not favor this treatment

2) ACTION/CAUTION:
- Active: Treatment requires action within weeks
- Cautious: treatment gives me months or longer to decide

3) SURGERY:
- No cutting: treatment does not require any surgery
- Cutting: surgery with some risks and hospital time

4) SEX:
- Sex: same as before treatment
- Sex: decreased compared to before treatment
- Sex: unable to engage in sex

5) URINARY:
- Urinary: no problems
- Urinary: short-term issues
- Urinary: long-term issues

6) BOWEL:
- Bowel: no problems
- Bowel: short term urgent & frequent bowel movements

7) LIFESPAN:
- Lifespan: live my expected lifespan

[12] Healthcare researchers have commented that learning which outcomes patients want to avoid is perhaps even more important than learning what they want, so the best-worst approach could have additional appeal.

Lifespan: live 5 years fewer than expected

The hybrid method we propose also overcomes weaknesses in (or objections to) the Best-Worst Case 2 approach, such as:

- It cannot accommodate attribute interaction effects (which are sometimes significant and quite useful for improving predictions)
- It does not formally support a compositional rule and therefore subsequent market simulations

Our proposed hybrid approach overcomes these two weaknesses. It obtains common scaling of all attribute levels and leverages conjoint judgments so that it justifies a compositional rule and therefore more formally supports market simulations. It also supports estimation of interaction effects due to its incorporation of CBC format choice tasks.

Another side benefit of our hybrid approach is a possibly enhanced ability to detect unengaged respondents (we test this below). With CBC, respondents can speed through the interview by using a simple decision rule (such as always picking the lowest priced product, or always selecting their favorite brand) and obtain extremely high internal fit (RLH) scores. With best-worst scaling, however, it's not as easy to speed through and still obtain a high fit (Elder and Pan 2009). To be rewarded with a high fit statistic, they must answer using more than just a superficial decision rule, making judgments of best and worst levels across a series of product profiles that when arranged in a chain of direct and indirect comparisons among the items in the attribute list are seen as internally consistent. In addition, those best-worst judgments must also be found to be consistent with subsequent CBC choices to achieve an overall high fit statistic within our hybrid method.

To summarize, the hybrid choice technique fusing Best-Worst Case 2 and Best-Worst Case 3 (Best-Worst CBC) has the following potential benefits:

- Utility scores for all levels are estimated on a common scale, allowing for the direct comparison of utilities across all attributes.
- Supports a compositional rule, permitting conjoint-style predictions using a market simulator.
- Can accommodate interaction effects.
- Better ability than standard discrete choice (CBC) methods to identify unengaged respondents.
- Better ability than CBC to handle within-concept level prohibitions (discussed below)

Experimental Design:

To motivate our approach to experimental design for this hybrid method, we first consider theory from the original card-sort conjoint analysis. With card-sort conjoint, the number of profiles to show each respondent must be at least as many as the number of parameters to estimate (to support individual-level estimation under OLS). With our hybrid conjoint approach, all parameters are estimated on a common scale, meaning that only one of the parameters needs to be fixed in order for the model to be identified. Thus, the number of parameters to estimate is equal to the total number of levels in the study including the None parameter (if it exists) minus 1.

Consider a typical conjoint analysis study with 6 attributes each with 4 levels. There are 24 levels in the study plus one None parameter. After selecting one parameter as the reference (for each attribute) and deleting its column from the design matrix (to identify the model), the number of main-effect parameters to estimate is 18. Our hybrid method is not the same as traditional card-sort conjoint, but this gives us a starting point for thinking about design generation.

Within each respondent, we should strive for a high degree of level balance (each level within each attribute appears about an equal number of times) and orthogonality (each level appears with each other level from different attributes an equal number of times). Furthermore, within the Section 3 CBC-formatted choice sets, previous research suggests that main effects are optimally estimated if minimal overlap is used; whereas to support interaction effects it is better to include at least a modest degree of level overlap (Sawtooth Software 1998). Sawtooth Software's Complete Enumeration and Balanced Overlap experimental design methods achieve these goals. Complete Enumeration can be used to design the profiles shown one-at-a-time (Section 2), and Balanced Overlap can be used for designing the Best-Worst CBC tasks (Section 3) so they include a modest degree of level overlap within the sets. We suggest multiple versions (blocks) of the plan to reduce order and context effects, as well as to increase the precision of interaction effects. Sawtooth Software's CBC design methods assume each respondent should receive a unique version (block), though one could ask the designer to generate only 4 to 6 blocks if collecting the data via paper-and-pencil.

We've suggested that Sawtooth Software's Complete Enumeration and Balanced Overlap design approaches could be used to generate the profiles for our hybrid method, but we should discuss how many profiles (cards) are needed for each respondent.

Let's first consider Section 2: the Best-Worst Case 2 section of the hybrid interview (the best-worst judgments of levels within individual profiles). With MaxDiff analysis (Best-Worst Case 1), we have found that stable individual-level estimates may be obtained under HB estimation if each item appears at least 3 times (Orme 2005). Considering the case of six attributes each with four levels, we would be showing a full-profile card each time (consisting of one level from each attribute), so it takes four cards shown to the respondent for each level to appear once. It would take 12 cards for each level to appear three times (the attribute with the largest number of levels is the constraint). However, we know that additional information is available in Sections 1 and 3 for stabilizing the data. Plus, we have evidence from the synthetic data set (Exhibit 2) that Best-Worst Case 2 doesn't require as many tasks relative to CBC. Therefore, it would seem unnecessary to ask each respondent to complete 12 cards in this case. With 8 cards, each item would appear two times; but results shown in Exhibit 2 suggests that may be excessive (especially given that information comes from the other two sections for further stabilizing results). Even so, we should at least obtain connectivity[13]among the items within Section 2. If each item appears just once, then there is a problem of lack of connectivity across the items, though the other two sections would provide enough additional information to ensure connectivity.

Lastly, we consider the Best-Worst CBC tasks (Section 3). Researchers have found that about 10 to 12 CBC tasks lead to quite adequate estimation of individual-level parameters when using HB for typically-sized CBC studies. Since we have additional information coming from the other two sections, it would therefore seem unnecessary to ask 10 tasks within this section. Probably around five would be adequate (we could test this with synthetic data simulations). If we show four concepts per task, five tasks would involve 20 total product concepts.

Consider again the hypothetical case of 6 attributes each with 4 levels. Based on previous research, our synthetic data set results (Exhibit 2), and rules of thumb for MaxDiff and CBC methodologies, we could

[13] Within MaxDiff experiments, connectivity means that all items are either directly or indirectly (via the law of transitivity) compared with all other items, allowing the researcher to place all items on a common scale.

make an educated approximation regarding an appropriate number of profiles to show in Section 2, the Best-Worst Case 2 section of our hybrid survey (6 profiles), and the Best-Worst CBC portion of the survey (Section 3: five tasks of four concepts each), for a total of 26 profiles (product concepts). If we use the Complete Enumeration experimental designer to generate the first 6 profiles (e.g. two tasks with three concepts each) and then use the Balanced Overlap experimental designer to generate the Best-Worst CBC tasks for the survey, then we should have a fine experimental design indeed.

In review, we would suggest the following rules of thumb for selecting the number of profiles and tasks to use in the hybrid questionnaire:

- **Section 2-** Best-Worst Case 2 tasks: If we weren't worried about respondent fatigue or time, we could select enough tasks so that each level appears at least 2x. If an attribute has five levels or more, this would mean at least 10 profiles, which would probably be excessive (given that additional information is available from the other two sections). A good rule of thumb would be to choose about 1.5K profiles, where K is the average number of levels in your study.

- **Section 3-** Best-Worst CBC tasks: Select five or six CBC-looking tasks, each with preferably 3 to 5 concepts per task.

With six attributes and four levels each, these rules of thumb might recommend 6 Best-Worst Case 2 profiles (Section 2) followed by 5 CBC-looking tasks (Section 3).

We should note that researchers employing ample sample sizes may decide that some precision at the individual-level could be sacrificed, and fewer choice tasks could be selected than the guidelines suggested above. For example, 5 Best-Worst Case 2 profiles could be shown (each item appears 1.25x) followed by three or four Best-Worst CBC tasks. Given the information also provided by the Priors section (Section 1), this should still be adequate information to use HB to estimate useful individual-level utilities.

It is also worth noting that enough information could be collected with our hybrid method (if the number of tasks and profiles is generous) to support purely individual-level estimation, via methods such as individual-level logit analysis.

Prohibitions

There are times when the researcher wishes to prohibit a level or levels of one attribute from appearing with a level or levels of another attribute (to avoid product combinations that are utterly and obviously impossible). Generally, this is discouraged as it leads to correlations between factors within the design matrix and thus reduced precision of parameter estimates. There are better ways for dealing with this. Alternative-specific CBC experiments can manage a significant number of prohibitions without necessarily leading to correlation within the design matrix. Attributes can be made conditional (specific) to certain alternatives, allowing one to customize the attribute list under any alternative. The classic example is when comparing cars, buses and trains: the attributes modifying cars are different from buses and trains. Such designs are more complex to set up, but demonstrate the power and flexibility of the CBC approach. In contrast, Best-Worst Case 2 (by itself) does not support alternative-specific designs, though a hybrid choice interview fusing Best-Worst Case 2 and Best-Worst CBC questions (alternative-specific design) could conceivably do so[14]. In any case, Best-Worst Case 2 (Section 2) as an element of the hybrid approach combining Priors (Section 1), Best-Worst Case 2 (Section 2), and CBC tasks (Section 3), is much more robust in the face of prohibitions than standard (generic attribute) conjoint analysis.

With traditional (generic) conjoint analysis or CBC involving a common attribute list for all alternatives, any prohibition between levels of different factors leads to correlation within the design matrix. However, two of the three sections of the hybrid approach are free from such intercorrelations. The pairwise choice tasks resulting from implied inequalities (for levels within the same attributes) resulting from the Priors section (Section 1) are not negatively affected by prohibitions. The Best-Worst Case 2 data (choices of best and worst levels within profiles—Section 2) don't lead to correlation in the design matrix when prohibitions are in place. Each row in the design matrix for Section 2 encodes the presence of a single attribute level (rather than conjoined sets of multiple attribute levels), so correlation structure is avoided. Only Section 3 (the Best-Worst CBC tasks) is affected by prohibitions in terms of introducing correlation within the design matrix. Thus, most of the observations in the design matrix are essentially immune to the negative effect of prohibitions on collinearity, significantly reducing the correlation within the overall design matrix. The effect of prohibitions on the hybrid approach can be simulated by generated random respondent data (respondents answering randomly), estimating aggregate MNL, and examining the standard errors of the parameters. Designs with and without prohibitions could be directly compared in this way, so the researcher could estimate the effect of prohibitions on the precision of the estimates prior to collecting real data.

We have suggested that the hybrid approach may be easier for less experienced researchers to use than CBC, due to the ease of interpreting the part-worth utilities. The robustness of the hybrid approach in the face of prohibitions is also a benefit that reduces the pitfalls and thus could lead to more confident usage by a wider group of market research analysts. Very shortly below we will discuss results of an empirical experiment where we test the hybrid approach versus the CBC standard.

Depth of Choice for Best-Worst Case 2 Tasks (Section 2):

In Section 2, we can ask the respondent to indicate the best and worst levels within each product profile. Exhibit 2 shows the potential gains for probing deeper to obtain additional best-worst choices within each profile. If we were studying 7 attributes, it would seem to make good sense to ask for best and worst levels, then remove those levels from the profile, and ask for best and worst again within the remaining 5 levels in the profile[15]. This ensures that we obtain more information regarding the levels of

[14] It is possible to consider an alternative-specific design for the hybrid approach. The CBC portion of the interview (Section 3) could be made to be alternative-specific. The Priors (Section 1) could provide rank-order information for levels within attributes. Best-worst choices of levels within alternative-specific profiles (Section 2) could be done: for example, best and worst levels within a bus profile; then best and worst levels within a train profile. The results of all three sections could be fused within a single MNL estimation, following the instructions in the Appendix. Note, however, that the independent variable matrix should be coded so that ASCs capture the relative preference of the label for the alternatives (i.e. bus, train, car) and attribute levels within each alternative are placed on a common scale. Thus, direct comparisons of utilities across attributes would only be supported within the attribute lists under *same* alternative (e.g. within bus levels only).

[15] The choice tasks should be coded to represent the full number of items in the first probe of best and worst levels and the reduced set of items available in the second probe of best and worst levels.

middling preference for placing all the attribute levels on a common scale. With just two attributes in the study, only two levels are shown in any one profile and we only need to ask the respondent to indicate which of the two is the level that makes him most want to buy the product. For studies with five or more attributes, it seems reasonable to ask for additional selections of best and worst beyond the first ones within each product concept.

An Experiment with Real Respondents to Compare Best-Worst Case 2, Best-Worst Case 3, CBC, and a Fusion of the Two Best-Worst Approaches:

In December 2012, we fielded a split sample experiment that we describe in further detail below. The main issues we wished to investigate were:

1. The similarity of Best-Worst Case 2 and CBC part-worths. Related to that, the ability of Best-Worst Case 2 and CBC to predict CBC-looking holdouts.
2. Whether fusing Best-Worst Case 2 and Best-Worst Case 3 (also known as Best-Worst CBC) would work better than Best-Worst Case 2 alone in predicting CBC-looking holdouts.
3. Respondent reaction to Best-Worst Case 2 compared to CBC (e.g. did respondents find the tasks confusing?) as well as time to complete the tasks.

Description of Split-Sample Experiment:
We used Survey Sampling International's (SSI) internet panel to target respondents in the upper three quartiles of household income and based on intention to purchase an HDTV within the next 12 months. We constructed an attribute list for HDTVs with 8 attributes and 25 total levels. Using Sawtooth Software's SSI Web platform, we interviewed approximately 1200 respondents during the 2nd week of December, 2012, randomly assigning respondents into four different questionnaire versions (design cells), as detailed below. (Screenshots from various sections of the survey are shown in Appendix B.)

Exhibit 4

Cell Name:	Sample Size:	Description:
Cell 1 (Holdouts)	329	• Screener Questions • Prior Ratings for attribute levels (without known order), 3-point scale + "no opinion" choice • 12 holdout CBC tasks (1 fixed version/block), 4 concepts/ task, high level overlap and utility balance, no "None", first choice only
Cell 2 (CBC)	304	• Screener Questions • Prior Ratings for attribute levels (without known order), 3-point scale + "no opinion" choice • 12 CBC tasks (randomized plan, moderate level overlap using *Balanced Overlap*), 4 concepts/ task, no "None", first choice only
Cell 3 (Hybrid, BW Case 2 FP + BW Case 3)	307	• Screener Questions • Prior Ratings for attribute levels (without known order), 3-point scale + "no opinion" choice • 6 Full-Profile Best-Worst Case 2 tasks, 8 items per set (profile), four clicks per task (1st best, 1st worst, 2nd best, 2nd worst), randomized level-balanced, near-orthogonal plan • 4 Best-Worst CBC (Best-Worst Case 3) tasks, 2 clicks per task (best concept, worst concept), randomized plan, moderate level overlap using *Balanced Overlap*), 4 concepts/task, no "None" • 4 holdout CBC tasks (1 fixed version/block), 4 concepts/task, high level overlap and utility balance, no "None", first choice only
Cell 4 (Hybrid, BW Case 2 PP + BW Case 3)	292	• Screener Questions • Prior Ratings for attribute levels (without known order), 3-point scale + "no opinion" choice • 12 Partial-Profile Best-Worst Case 2 tasks, 5 items per set (profile), two clicks per task (best and worst), randomized level-balanced, near-orthogonal plan • 4 Best-Worst CBC (Best-Worst Case 3) tasks, 2 clicks per task (best concept, worst concept), randomized plan, moderate level overlap using *Balanced Overlap*), 4 concepts/task, no "None" • 4 holdout CBC tasks (1 fixed version/block), 4 concepts/task, high level overlap and utility balance, no "None", first choice only

(Notes: The only difference between Cells 3 and 4 is underlined above. Qualitative questions regarding respondent experience with the survey were also asked within each of the four cells.)

Median Time to Complete Selected Sections:

Exhibit 5

Cell 1:	
Prior Ratings Grid for Levels:	35 seconds (0.58 min)
12 CBC tasks, 4 concepts each, first-choice only:	177 seconds (2.95 min)
Total:	**212 seconds (3.5 min)**
Cell 2:	
Prior Ratings Grid for Levels:	40 seconds (0.67 min)
12 CBC tasks, 4 concepts each, first-choice only:	170 seconds (2.83 min)
Total:	**210 seconds (3.5 min)**
Cell 3:	
Prior Ratings Grid for Levels:	36 seconds (0.60 min)
6 Best-Worst Tasks (Full-Profile) (4 clicks: 1st best, 1st worst, 2nd best 2nd worst):	192 seconds (3.20 min)
4 Best-Worst CBC Tasks:	79 seconds (1.32 min)
Total:	**307 seconds (5.1 min)**
Cell 4:	
Prior Ratings Grid for Levels:	39 seconds (0.65 min)
12 Best-Worst Tasks (Partial-Profile) (2 clicks: best and worst):	186 seconds (3.10 min)
4 Best-Worst CBC Tasks:	83 seconds (1.38 min)
Total:	**308 seconds (5.1 min)**

What is striking about the timing results above is how quickly/efficiently respondents can do these complex tasks. Granted, the attribute levels were very concise, but there are eight attributes to describe each HDTV in full profile. Respondents completed CBC tasks requiring just first choices in an average of about 15 seconds per task. Best-Worst CBC tasks (CBC tasks requiring choice of best and worst concepts per set) required 20 seconds per task (an extra 5 seconds to capture the second choice). Best-Worst Case 2 tasks (choice of best and worst level within a single concept) required 8 seconds per click. The Priors rating grid just took about 40 seconds (about 3 seconds to rate each attribute level). In sum, the standard CBC questionnaire (including the Priors section) took just 3.5 minutes, and the hybrid questionnaire (with its three sections) took just 5.1 minutes. For an 8-attribute, 25-level study, this does not seem very time consuming.

Qualitative Assessment of Survey:

After either the CBC or Best-Worst Case 2 questions, respondents evaluated their experience with that section of the survey (listed as variables Qual1 through Qual5 in Exhibit 6). At the end of the entire questionnaire, respondents evaluated the entire survey experience (Qual6). In all cases we used a 5-point Likert scale (1=strongly disagree, 5=strongly agree). Mean scores (with standard errors in parentheses) are shown. The evaluations were nearly identical in most every case. We ran six separate F tests (one for each of the Qual variables in Exhibit 6). The only statistically significant differences among the design cells (95% confidence) were for variables Qual2 and Qual3. Although these are statistically significant differences, given how similar the mean ratings are, they do not represent very meaningful differences in the way respondents perceived these different conjoint and MaxDiff questionnaires.

Exhibit 6

	Cell 1 **CBC** **N=329**	**Cell 2** **CBC** **N=304**	**Cell 3** **BW2FP** **N=307**	**Cell 4** **BW2PP** **N=292**
Qual1: This section was at times monotonous and boring	2.36 (0.069)	2.39 (0.070)	2.36 (0.682)	2.38 (0.071)
Qual2: I'd be interested in taking another survey that included a section like this in the future *	4.33 (0.050)	4.27 (0.054)	4.28 (0.051)	4.13 (0.060)
Qual3: The format of the questions in this section made it easy for me to give realistic answers that reflect what I'd do if buying a real HDTV *	4.16 (0.052)	4.07 (0.057)	4.03 (0.055)	3.91 (0.059)
Qual4: The way the HDTVs were presented made me want to slow down and make careful choices	4.05 (0.053)	4.00 (0.050)	4.01 (0.052)	3.96 (0.054)
Qual5: This section was confusing	1.86 (0.065)	1.89 (0.064)	1.90 (0.067)	2.01 (0.069)
Qual6: (Rating of the overall survey, after completion of all parts of the survey)	4.16 (0.043)	4.08 (0.047)	4.16 (0.046)	4.08 (0.044)
Total Survey Time (Median)	6.2 min	7.8 min	9.5 min	9.4 min

(* indicates significant difference among the groups based on F Test, 95% confidence. For variables Qual3 and Qual4, additional independent t-tests between respondent groups showed that Cell 1 had a higher mean than Cell 4 at the 95% confidence level.)

Exhibit 7

Aggregate Logit Results

(After Zero-Centering Effects within Attributes for Comparison)

		CBC	BW2FP	BW2PP
		Cell 2	Cell 3	Cell 4
		n=304	n=307	n=292
Brand:				
	Panasonic	-0.16	-0.21	-0.08
	Samsung	0.23	0.35	0.50
	Sharp	-0.09	-0.39	-0.55
	Sony	0.19	0.74	0.75
	Visio	-0.15	-0.56	-0.68
	LG	-0.03	0.07	0.06
Screen Size:				
	32 inches	-0.71	-1.42	-1.26
	46 inches	0.18	0.50	0.28
	55 inches	0.53	0.92	0.98
Technology:				
	Plasma	-0.24	-0.56	-0.56
	LED Backlit LCD	0.24	0.56	0.56
Contrast:				
	Superior contrast	0.13	0.32	0.37
	Good contrast	-0.13	-0.32	-0.37
Warranty:				
	1 year	-0.11	-0.96	-0.94
	2 year	0.03	0.14	0.09
	3 year	0.08	0.82	0.85
Internet Connectivity:				
	None	-0.23	-1.18	-1.21
	Wired	-0.02	0.01	-0.03
	Wired & Wi-Fi	0.25	1.17	1.24
3D Support:				
	Supports 3D	0.14	0.44	0.52
	No 3D support	-0.14	-0.44	-0.52
Price:				
	$500	0.44	1.10	1.25
	$750	0.16	0.41	0.39
	$1,000	-0.19	-0.66	-0.57
	$1,250	-0.41	-0.85	-1.06
Standard Deviation:		0.27	0.72	0.75

One of the main benefits of using Best-Worst Case 2 is that the resulting parameter estimates are all placed on a common scale (Exhibit 14). But, to be able to compare the parameters to CBC, we need to zero-center the utility scores within each attribute, which we've done above in Exhibit 7. To summarize the magnitude of the parameters, we've computed the standard deviation of each column.

The utilities for the two Best-Worst Case 2 cells are more than double the size of the CBC part-worths. This suggests respondents answer Best-Worst Case 2 questionnaires with about half the response error as CBC.

We also ran a simple correlation analysis among the three columns of part-worth utility scores:

Exhibit 8
Correlations among Utility Scores

	Cell 2 CBC	Cell 3 BW2FP	Cell 4 BW2PP
Cell 2 CBC	1.00		
Cell 3 BW2FP	0.91	1.00	
Cell 4 BW2PP	0.90	0.99	1.00

The CBC utilities are correlated 0.91 and 0.90 with the Cell 3 and Cell 4 Best-Worst Case 2 utilities, showing high correspondence. The two sets of Best-Worst Case 2 utilities are correlated at 0.99, demonstrating what appear to be extremely similar results between the two Best-Worst Case 2 approaches (full-profile showing 8 levels per profile, with drill-down to best 2 levels and worst 2 levels, vs. partial-profile showing 5 levels per profile).

The Swait-Louviere test allows us to compare more formally the results of the aggregate parameters. We compared Cell2 (CBC) to Cell 3 (BW2FP), Cell2 (CBC) to Cell 4 (BW2PP), and Cell 3 (BW2FP) to Cell 4 (BW2PP). Each of the three tests for parameter equivalence between the cells rejects the null hypothesis at better than 99.9% confidence. We are highly confident that the aggregate logit parameters for each of the cells are different from the other.

We were also interested in the relative design efficiency for the two different Best-Worst Case 2 questionnaires. Cell 3 involved full-profiles, asking respondents to select best and worst levels among all 8 attribute levels describing an HDTV profile. Then, those two choices were taken away, and respondents were asked to select next-best and next-worst levels among the remaining attributes. Cell 4 involved a more traditional best-worst approach, where five attribute levels were shown per profile (avoiding the selection of any two levels from within the same attribute) and respondents were asked to pick the one best and one worst level per set. Both experiments involved 24 total clicks, since six full profiles were shown for Cell 3 (4 clicks per task) and 12 partial-profile sets were shown for Cell 4 respondents (2 clicks per task). It could be argued that the Cell 3 design was at an advantage in terms of statistical efficiency, since more items overall were shown (sets of either 8 items or 6 items to evaluate, compared to sets of 5 items as shown in the partial-profile sets). But, the partial-profile tasks had the opportunity to do a better job of balancing the number of exposures for each attribute level. For example, each level within a 2-level attribute (such as 3D Capability) would appear three times as often as each level within a 6-level attribute (such as Brand) within the full-profile cell. With the partial-profile cell, the exposures of each level can be nearly perfectly balanced, since for example the Brand attribute can be oversampled compared to the 3D Capability attribute. It turned out that the standard errors from the aggregate logit estimation were slightly smaller for the partial-profile Best-Worst Case 2 cell. The geometric means of the standard errors (across all 24 parameters) were 0.098904 and 0.101210 for Cells 4 and Cells 3 respectively. By taking the ratio of the squares of these standard deviations, we compute that the partial-profile Best-Worst Case 2 design was 5% more efficient than the full-profile Best-Worst Case 2 design. Since both Best-Worst Case 2 sections took almost exactly the same amount of time to complete, this would seem to very slightly favor the partial-profile approach in terms of statistical efficiency per comparable unit of respondent effort.

Hierarchical Bayesian Analysis of Utility Scores

We estimated individual-level utilities using Sawtooth Software's CBC/HB. To help stabilize the individual-level estimates, we incorporated within-attribute level information from the Priors ratings grid as ordinal information within the utility estimation (soft constraints, as suggested by Lattery 2009)[16]. This involves formatting a choice task for implied paired comparisons within attributes. For example, if the respondent rated (using the 3-point scale) screen sizes in the following order: 46-inch > 55-inch > 32-inch, then three additional choice tasks were formatted within the data file, indicating that 46-inch was preferred to 55-inch, 46-inch was preferred to 32-inch, and 55-inch was preferred to 32-inch. These are formatted as extreme partial-profile tasks, where attributes not involved in the paired comparison are retained at zeros within the independent variable matrix. If two levels were rated equally or if the respondent clicked that he had no opinion, then the choice task representing the implied paired comparison was dropped (in CBC/HB data files, tasks where the answer column is left as zeros for all alternatives in the task are automatically dropped).

A prior variance of 1 was used, applying a "proper prior covariance" coding as suggested by Lenk (Lenk 2009), with 5 degrees of freedom. 10,000 initial iterations were used, followed by 10,000 used iterations.

In addition to fusing the rank-order information from the Priors rating task (Section 1), we also fused Best-Worst Case 2 and Best-Worst Case 3 tasks (Sections 2 and 3) within the same utility runs. No attempt was made to adjust for differences in scale (response error) between the different sections of the questionnaire.

We computed seven different utility runs. Each of the HB runs took no more than 12 minutes, using a standard Lenovo laptop computer with only a modest amount of computing power per today's standards. The part-worth utility scores (after zero-centering within factor to allow comparisons and using the method of within-respondent scale standardization called *zero-centered diffs*) are shown below. We've highlighted columns representing Best-Worst Case 2 only information (non-conjoint data) to facilitate visual comparisons between part-worth scores using non-conjoint techniques (Best-Worst Case 2) versus conjoint techniques (CBC and Best-Worst Case 3).

[16] We should note that soft constraints can lead to biasing attributes with a lot of levels to have more impact on product choice than attributes with fewer levels. A more defensible approach that should reduce this potential bias is to implement monotonicity (utility) constraints at the individual level via a method such as HB (as is implemented in ACA/HB or even ACBC/HB when applying customized utility constraints).

Exhibit 9

	Cell2_Priors+ CBC	Cell3_Priors+ BW2	Cell3_Priors+ BW3	Cell3_Priors+ BW2+BW3	Cell4_priors+ BW2	Cell4_priors+ BW3	Cell4_Priors+ BW2+BW3
	n=304	n=307	n=307	n=307	n=292	n=292	n=292
Brand:							
Panasonic	-21.42	-15.83	-8.12	-11.50	-8.32	-12.08	-8.02
Samsung	30.49	18.85	13.95	18.65	23.35	25.59	24.68
Sharp	-15.67	-23.82	-13.28	-18.35	-25.58	-23.69	-26.94
Sony	28.04	37.55	28.97	35.14	33.55	32.02	34.84
Visio	-20.68	-24.73	-25.27	-28.39	-27.16	-22.77	-27.61
LG	-0.76	7.98	3.76	4.45	4.17	0.93	3.04
Screen Size:							
32 inches	-95.70	-64.32	-88.86	-76.87	-50.19	-79.42	-67.29
46 inches	27.36	20.64	26.80	22.85	12.38	19.03	14.13
55 inches	68.34	43.68	62.07	54.02	37.81	60.40	53.16
Technology:							
Plasma	-33.09	-22.89	-29.71	-24.78	-21.69	-22.64	-18.28
LED Backlit LCD	33.09	22.89	29.71	24.78	21.69	22.64	18.28
Contrast:							
Superior Contrast	22.55	22.57	19.21	17.89	26.20	22.18	21.34
Good Contrast	-22.55	-22.57	-19.21	-17.89	-26.20	-22.18	-21.34
Warranty:							
1 year	-26.97	-53.57	-33.70	-41.63	-52.59	-35.54	-41.13
2 year	1.38	4.72	1.14	3.02	2.42	1.45	2.39
3 year	25.59	48.86	32.56	38.61	50.17	34.09	38.74
Internet Connectivity:							
None	-39.44	-60.69	-47.21	-54.57	-61.17	-46.54	-54.13
Wired	-0.62	0.20	3.85	2.57	-1.58	-0.45	0.31
Wired & Wi-Fi	40.06	60.49	43.36	51.99	62.75	47.00	53.81
3D Support:							
Supports 3D	14.61	17.71	12.51	17.61	22.34	18.53	21.64
No 3D support	-14.61	-17.71	-12.51	-17.61	-22.34	-18.53	-21.64
Price:							
$500	68.13	66.34	78.70	67.73	67.42	72.49	65.28
$750	24.31	20.94	28.09	23.90	18.33	26.97	22.68
$1,000	-17.68	-24.69	-19.97	-21.66	-23.45	-18.43	-20.02
$1,250	-74.76	-62.59	-86.82	-69.98	-62.31	-81.03	-67.94

Exhibit 10

	Cell2_Priors+ CBC	Cell3_Priors+ BW2	Cell3_Priors+ BW3	Cell3_Priors+ BW2+BW3	Cell4_priors+ BW2	Cell4_priors+ BW3	Cell4_Priors+ BW2+BW3
	n=304	n=307	n=307	n=307	n=292	n=292	n=292
Average Importances	Total	Total	Total	Total	Total	Total	Total
Brand	17.02	16.61	17.43	17.81	14.98	16.94	16.61
Screen Size	23.18	16.74	21.62	19.38	15.42	21.38	19.28
Technology	10.53	9.31	9.50	8.86	9.58	9.29	8.82
Contrast	5.89	5.71	4.93	5.13	6.55	5.63	5.57
Warranty	7.15	12.82	8.35	10.28	12.86	8.76	10.46
Internet Connectivity	10.67	15.27	11.42	13.71	15.59	11.94	14.16
3D Support	5.71	7.29	5.66	6.84	8.63	6.45	7.56
Price	19.85	16.25	21.09	17.99	16.38	19.63	17.54

Correlations among Utility Scores:

Exhibit 11

	Cell2 (Priors +CBC)	Cell3 (Priors+ BW2FP)	Cell3 (Priors +BW3)	Cell3 (Priors+ BW2FP+ BW3)	Cell4 (Priors +BW2PP)	Cell4 (Priors+ BW3)	Cell4 (Priors+ BW2PP+ BW3)
Cell2 (Priors+CBC)	1.00						
Cell3 (Priors+BW2FP)	.94	1.00					
Cell3 (Priors+BW3)	.99	.95	1.00				
Cell3 (Priors+BW2FP+BW3)	.97	.99	.99	1.00			
Cell4 (Priors+BW2PP)	.91	.99	.93	.97	1.00		
Cell4 (Priors+BW3)	.98	.97	.99	.99	.96	1.00	
Cell4 (Priors+BW2PP+BW3)	.96	.99	.97	.99	.98	.99	1.00

There is a lot of data here, so we summarize with the following observations:

1. As seen with the aggregate logit analysis, the two separate groups of respondents who received different Best-Worst Case 2 questionnaires (full-profile vs. partial profile) result in nearly identical scores (correlation of 0.99).
2. Best-Worst Case 2 utilities and CBC utilities are highly correlated, at 0.91 and 0.94, for the two different versions of Best-Worst Case 2. This is similar to the findings from aggregate logit, where the correlations were 0.90 and 0.91.
3. Best-Worst Case 2 parameters seem less similar to CBC utilities than Best-Worst Case 3 utilities are to CBC utilities, as we'd expect, given the method variance differences.

Correlations among Shares of Preference for CBC Scenarios (12 Tasks x 4 Concepts Each)
The common out-of-sample predictive test checks how well the part worths from one set of respondents can predict the choices of a different set of respondents. Cell 1 contained 329 respondents who each completed the same 12 CBC-looking holdout tasks (1 fixed version/block). Each task contained 4 concepts. We constructed these tasks to be especially difficult, by starting with a purely randomized design (involving a very large amount of level overlap), and then manually utility balancing the tasks even further so that each concept had about as many positive aspects and negative aspects as the other concepts within the same task. This resulted in 48 total product concepts (sets of 4 times twelve tasks). Across the 329 respondents, we tallied up the percent of choices among the concepts. This gave us a vector of 48 shares of choice to predict using market simulators based on the utility scores from the other three groups of respondents.

We used the HB utilities estimated using various components of information (depending on the design cells and treatments) within a market simulator, predicting out-of-sample shares of choice for the Cell 1 respondent CBC holdout scenarios with Sawtooth Software's Randomized First Choice method[17]. We tuned the scale factor for each separate simulation run so that the standard deviation across the 48 aggregated product shares was approximately equal to the standard deviation of the observed holdout shares. Then, we computed the correlation between the predicted and holdout shares for seven different utility estimations. This is shown in the grey column below. We also computed the correlations among the predictions across all estimated sets of utilities. (The patterns of these results are nearly identical to the previous correlations comparing the average vectors of utilities in Exhibit 11.)

Exhibit 12

	Cell1 (Holdouts)	Cell2 (Priors +CBC)	Cell3 (Priors+ BW2FP)	Cell3 (Priors +BW3)	Cell3 (Priors+ BW2FP+ BW3)	Cell4 (Priors +BW2PP)	Cell4 (Priors+ BW3)	Cell4 (Priors+ BW2PP+ BW3)
Cell1Holdouts	1.00							
Cell2 (Priors+CBC)	.91	1.00						
Cell3 (Priors+BW2FP)	.85	.94	1.00					
Cell3 (Priors+BW3)	.86	.98	.95	1.00				
Cell3 (Priors+BW2FP+BW3)	.87	.97	.99	.98	1.00			
Cell4 (Priors+BW2PP)	.83	.93	.99	.94	.98	1.00		
Cell4 (Priors+BW3)	.90	.99	.97	.99	.99	.97	1.00	
Cell4 (Priors+BW2PP+BW3)	.88	.96	.99	.96	.99	.99	.99	1.00

The best out-of-sample prediction of the holdout choices (correlation of 0.91) comes from Cell 2 (Priors + CBC). This shouldn't surprise us, since the holdout CBC tasks are identically formatted as the Cell 2 CBC tasks. The next highest prediction accuracy comes from Cell 4's Priors + Best-Worst Case 3 (Best-Worst CBC) tasks. Best-Worst CBC tasks are also identically formatted as the holdout tasks, except both best and worst concepts are selected in each Task. The hybrid fused models[18] involving Priors + Best-

[17] We also checked the performance of the standard logit simulation method (Share of Preference), finding RFC slightly outperforming it for predicting these holdouts for this data set.

[18] As we noted previously, we haven't taken any additional steps during utility estimation to adjust for scale differences among the different choice contexts within the fused datasets. Perhaps extensions that adjust for scale differences could achieve slightly better results. Also, it is possible to adjust the contribution of the Best-Worst Case 2 and Best-Worst Case 3 sections (sections 2 and 3), by selecting more or fewer choice tasks for respondents to complete of each type. Also, reweighting between the sections could be done by changing the

Worst 2 + Best-Worst 3 tasks (sections 1 through 3 of the hybrid questionnaire)) achieved prediction accuracies of 0.87 and 0.88 for Cells 3 and 4. Including Best-Worst Case 3 tasks (section 1 and section 3) within the fusion performed better than when only Best-Worst Case 2 tasks (section 1 and section 2) were involved.

Within-Sample Hit Rates for CBC Scenarios (4 Tasks x 4 Concepts Each)

In addition to out-of-sample predictions, we also included four fixed holdout CBC tasks within each cell of our experiment. This allows us to conduct the common within-sample hit-rate tests. For each respondent, we sum the utilities for levels comprising the four concepts involved in a holdout choice task. The concept with the highest utility is predicted to be chosen. If the prediction matches what the respondent actually selected, this is counted as a hit. We constructed the holdout CBC tasks to be very difficult to predict. We used a high degree of level overlap, and manually constructed the tasks so that no concept would dominate another. In one of the holdout tasks, we held brand and price constant, only allowing the other six attributes to vary across the four concepts.

The hit rate accuracies for the different cells of the experiment and ways to estimate the utilities are shown below[19].

Exhibit 13

Hit Rate Accuracies

	Hit Rate
Cell2 (Priors+CBC)	.56
Cell3 (Priors+BW2FP)	.50
Cell3 (Priors+BW3)	.50
Cell3 (Priors+BW2FP+BW3)	.53
Cell4 (Priors+BW2PP)	.53
Cell4 (Priors+BW3)	.53
Cell4 (Priors+BW2PP+BW3)	.54

Cell 2 with the fusion of priors as soft constraints and CBC tasks achieved the highest hit rate, with 56% of holdout choices predicted accurately (no hard utility constraints were employed for any utility estimation within this paper). Again, this shouldn't surprise us since the CBC tasks used to estimate the utilities were formatted exactly like the subsequently answered CBC holdout tasks. The next most successful predictions came from Cell 4 with the hybrid fusion of sections 1, 2, and 3 (54%). Cell 3's hybrid fusion involving sections 1, 2 and 3 performed nearly as well, with 53% hit rate accuracy. Even the Priors + Best-Worst Case 2 fusion (sections 1 and 2, which isn't technically conjoint analysis at all) achieved commendable hit rate accuracies of holdout CBC tasks, with hit rates of 50% and 53% for Cells 3 and 4.

For Cell 2 (Priors + CBC) that achieved the best out-of-sample predictions and hit rates, we re-estimated the utilities using covariates within HB. The covariates included a screener question asking respondents how much they paid for their last TV and also a multi-select question wherein we asked respondents to check their top four (out of eight) attributes in terms of importance. The covariates run resulted in

magnitude of the X values in the design matrix for one section of the choice data relative to the other, or replicating certain tasks within the utility estimation matrix.

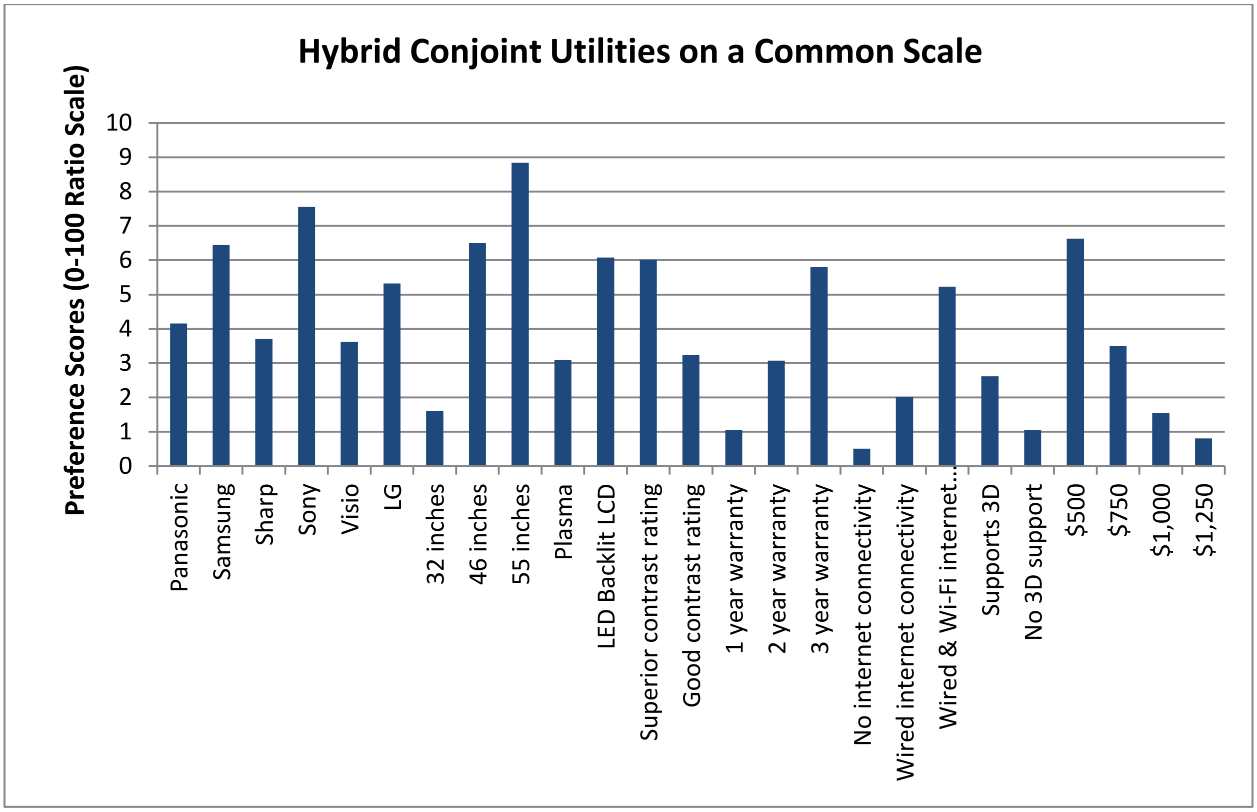
[19] Differences between predictions of at least 4% represent significant differences (95% confidence level).

almost exactly the same hit rate and holdout share prediction rate (except slightly lower). This replicates findings from previous researchers at the Sawtooth Software conferences who have not found covariates to improve the predictive accuracy of CBC.

Key Benefit of Best-Worst Case 2 Fusion: Scores on a Common Scale

Obtaining utility scores for attribute levels on a common scale (within *and* across attributes) is arguably the biggest benefit of using Best-Worst Case 2—alone or fused with Best-Worst Case 3 (Best-Worst CBC). We could just report the average raw HB utility scores for the sample, but these typically involve positive and negative values on an interval scale (not as easily interpreted as positive ratio-scale values). Also, these raw scores would not place each respondent on equal footing. Respondents who were more consistent would receive larger scale, giving them more weight in the sample mean computations than respondents who were less consistent. With Sawtooth Software's MaxDiff software, we use a 0-100 ratio scale exponential transformation that makes each respondent's sum of scores equal 100, and thus gives each respondent equal weight. The transformation is described in the MaxDiff software manual. We have used this same transform with the HB-estimated scores for the hybrid conjoint for Cell 3 (fusing sections 1, 2, and 3) and presented these 0-100 ratio scale scores that sum to 100 in Exhibit 14[20].

Exhibit 14

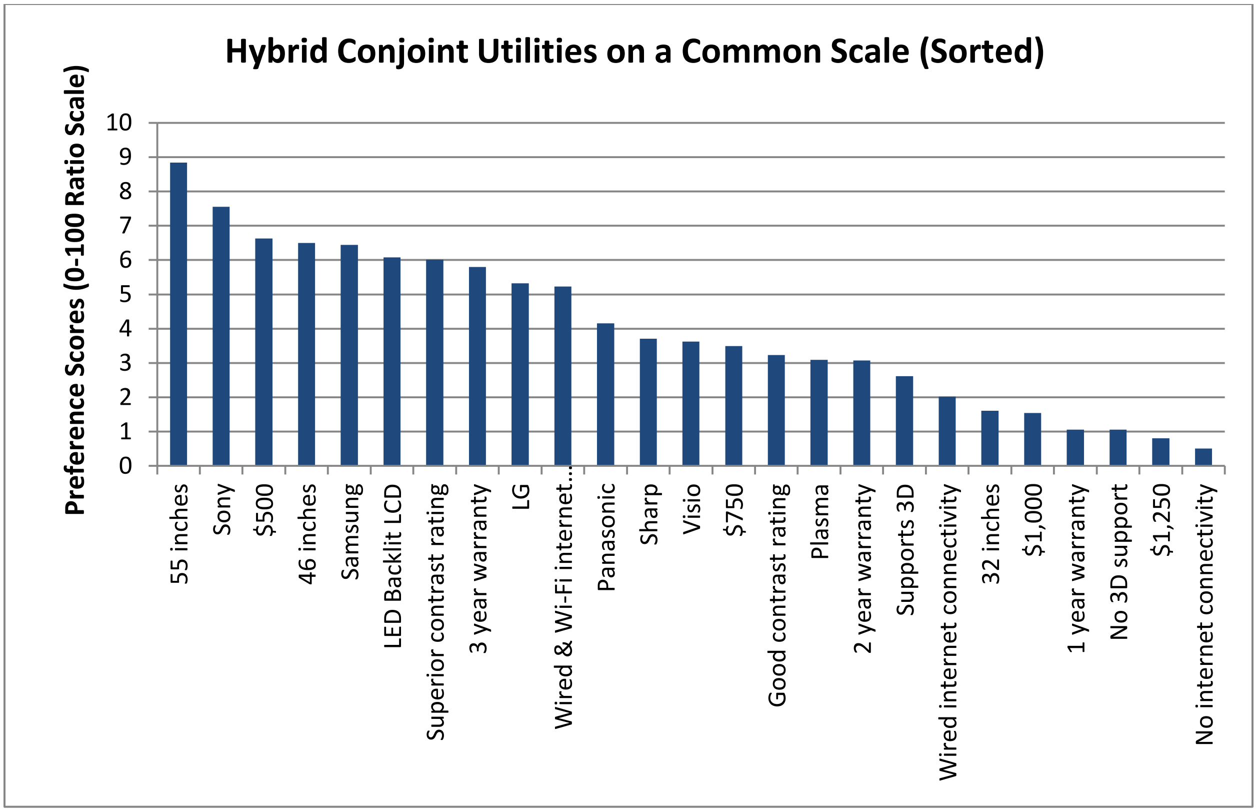


[20] We could have shown the utility scores for Best-Worst Case 2 separately from the scores resulting from the fusion of Best-Worst Case 2 and Best-Worst Case 3, but the average HB scores (leveraging the priors information, under the 24-parameter model) for the two are correlated at 0.99. So, we have shown only the results of the full hybrid fusion of choice information, since it contains more information and is the focus of this paper.

With standard conjoint analysis, it is not mathematically possible to compare one level of one attribute to one level of a different attribute. With Best-Worst Case 2 (or the fusion of it and either CBC or Best-Worst CBC), we may do so. The feature that respondents reported made them most want to buy the HDTV was “55-inch display” (a score of nearly 9). As an example of a comparison involving two separate attributes, both “LED Backlit LCD” and “Superior contrast rating” receive a score of 6, meaning that these two features are equally likely to be selected as features that make respondents want to purchase an HDTV.

We may prioritize the attribute levels, by rank-ordering them by the same mean scores presented in Exhibit 14. We show those rearranged results in Exhibit 15:

Exhibit 15



Among the attributes we included in the study, the three attribute levels that respondents state make them most want to buy an HDTV (on average) are “55 inches”, “Sony”, and $500”. The three attribute levels that make respondents least want to buy an HDTV are “No internet connectivity”, “$1,250”, and “No 3D support”. As one example of the kinds of comparisons that may be made when leveraging Best-Worst Case 2 data (and transforming to the ratio scale), the attribute “Superior contrast rating” (score of 6) is twice as likely to be picked as an attribute that makes respondents want to buy an HDTV compared to “2 year warranty” (score of 3). Having the scores all placed on a common ratio scale (all positive values) makes it very easy to interpret the results. The common pitfalls (lack of ability to compare utilities between attributes and lack of ratio scale) that often bedevil newcomers to conjoint analysis are avoided.

The ability to directly compare the utility scores across all levels (including those from different attributes) provides additional strategic information (beyond CBC) regarding how to motivate buyers (or patients, in the case of healthcare topics; or employees in the case of employment satisfaction research). Also, using scores on a common scale within segmentation methods (such as latent class or cluster analysis) leverages more information (more dimensions of preference) than standard conjoint scores (where the differences in utilities are only meaningful within attribute). There are 25 total levels and 8 attributes in this study. Under CBC, since levels within each attribute are scaled with respect to an arbitrary constant, there are only 25-8 = 17 degrees of freedom (parameters to estimate, or independent bits of preference information). Under Best-Worst Case 2, there are 25-1 = 24 degrees of freedom (parameters to estimate, or independent bits of preference information). Thus, assuming that respondents can state reliably which levels are drivers and detractors of preference, Best-Worst Case 2 has the opportunity to provide more bits of preference information than CBC, ACBC, ACA, or card-sort conjoint (CVA). Whether the enhanced information leads to improved market segmentation discovery is an open question for future research. When deciding to segment using one method or the other, we shouldn't lose sight of the fact that (after zero-centering within attribute for direct comparison) Best-Worst Case 2 uncovers a statistically different utility vector than CBC, and CBC's utilities are more predictive of choices within CBC-looking holdout tasks.

Detecting Unengaged Respondents:

One of the less important aims of our research into a new hybrid choice method was to investigate whether the fit statistic from CBC/HB would be a better indicator of respondent engagement (quality) than the fit statistic from a CBC interview. With CBC, a respondent who was attempting to shortcut the interview by using a very simple decision rule (such as always picking the lowest price product) can be rewarded with a very high fit statistic. However, MaxDiff choices require thoughtful responses to achieve a high internal fit score for a respondent.

We included a question early in the questionnaire (prior to the conjoint questions) regarding which room the respondent would most likely put a new HDTV. There were nine options given: Den, Home office, Kitchen, Living room, Home theater room, Bedroom, Bonus room, Other place, and Don't know. The options were randomized, except for the last two. The same question was repeated near the end of the questionnaire (with options randomized again, leading to a different order), after the conjoint section. We compared the responses as a proxy for respondent engagement. It is possible that respondents could legitimately have changed their mind about where the new HDTV would go after having completed the conjoint section. So, we cannot assume mismatched answers convicts a respondent as unengaged or “bad”. That said, nine out of ten respondents demonstrated perfect consistency. This seems impressive, especially since we did no preliminary data cleaning to trim speeders or straightliners before tabulating these results. All respondents as sent to us from SSI panel were used.

We sorted respondents by their RLH statistic into three equal-sized buckets. Then, we calculated the test-retest reliability for each of the three segments (see Exhibit 16).

Exhibit 16

Test-Retest Reliability (In Which Room Would You Most Likely Put a New HDTV?)

	Cell 2 (CBC)	Cells 3 & 4 (BW2+BW3)
Lowest 1/3 RLH Fit	81.2%	80.9%
Middle 1/3 RLH Fit	96.0%	93.6%
Upper 1/3 RLH Fit	96.1%	98.1%
Average:	91.1%	90.9%

There seems to be only modest evidence that the RLH fit statistic from the hybrid methods (Cells 3 and 4) provides better discrimination on the test-retest reliability. But, such a result should be expected, since the hybrid methods require many more clicks than Cell 2 (the CBC questionnaire) and thus more opportunity to distinguish between consistent and inconsistent respondents.

Summary Observations and Conclusions:

Best-Worst Case 2 Provides Common Scaling for All Attribute Levels within a Conjoint-Type Experiment: The common scaling across all attribute levels is a unique benefit for Best-Worst Case 2. This could be especially valuable for working with less sophisticated clients and for certain types of research projects wherein it would be useful to be able to directly compare attribute levels across the attribute list (such as messaging, healthcare topics, and employment research).

Best-Worst Case 2 Questionnaires Are Efficient: From a statistical standpoint, fewer Best-Worst Case 2 tasks are needed relative to CBC tasks to provide equal information. More importantly, respondents answer Best-Worst Case 2 tasks with about half the error rate as standard CBC tasks. This means that Best-Worst Case 2 overall is a more efficient way to elicit preference data from respondents than CBC.

The Use of HB Enhances the Efficiency for either CBC or our proposed hybrid choice methodology. Stable individual-level estimates may be obtained for a moderately large design with 25 attribute levels, with tradeoff question sections taking the typical respondent less than six minutes, using HB estimation that takes less than 15 minutes for the analyst to run. We are becoming fond of using *a priori* and stated prior rankings (as soft constraints) to inform HB estimation of within-attribute rank order using appended pairwise comparison tasks.

Respondents Find CBC and Best-Worst Case 2 Questionnaires Equally Appealing and Clear/Realistic: The qualitative evaluation of the surveys showed little differences between reactions to CBC and Best-Worst Case 2 questionnaires.

Best-Worst Case 2 Is Not a Substitute for CBC: Although Best-Worst Case 2 results in utility parameters that are correlated with CBC at about 0.90 or better across multiple studies, the results are not equivalent (after controlling for scale). Our study shows that predictions of CBC-looking holdout tasks, though quite good, are not quite as high as for standard CBC. Thus, Best-Worst Case 2 is not a substitute for CBC. Its inventor, Jordan Louviere didn't claim it was. He positioned it more as a complement to CBC—providing greater strategic information due to its ability to place all attribute levels on a common utility scale.

Best-Worst Case 2 Involves Stated Rather than Derived Preferences: Socially desirable responses could bias Best-Worst Case 2 utilities. In contrast, conjoint analysis techniques derive the preference scores by observing choices of product wholes. Respondents may not be willing or able to reveal why they select or avoid product concepts; they just do it. Asking them to indicate which levels are most or least preferred may be counterproductive in terms of uncovering the true and deep-rooted reasons behind their choice behavior. Even if this is the case, one should not ignore the almost surprising success of Best-Worst Case 2 in predicting CBC-looking holdouts across multiple studies by independent researchers.

No Evidence to Favor Full-Profile over Partial-Profile Best-Worst Case 2: Both sets of parameters are correlated about 0.99. From a statistical efficiency standpoint, the partial-profile methodology holds a slight edge over the full-profile Best-Worst Case 2 questionnaire. Respondents completed either series of questions (both requiring 24 clicks) in about equal time. The parameters were not equivalent (per the Swait-Louviere test), but both full-profile and partial-profile Best-Worst Case 2 formats predicted CBC holdouts about equally well. Without more evidence, it seems both approaches are equally viable.

The Fusion of Best-Worst Case 2 and Best-Worst Case 3 Has Appeal: Fusing the two types of data means that the choice interview can be labeled a conjoint method (though a hybrid one). Including Best-Worst Case 3 tasks within the fusion leads to part-worths that are more predictive of CBC holdouts. But, the fusion of the two means that two slightly different types of choice tasks are being fit using a single utility vector, leading to a compromise between the two methodologies. The fusion (especially given the additional within-attribute prior ranking data) leads to strong and stable individual-level utilities with HB estimation, but it does not predict CBC holdouts better than fusing prior ranking data with CBC tasks alone. A nice side benefit is that fusing Best-Worst Case 2 and Case 3 makes it possible for researchers to model interaction effects, if they exist, whereas Best-Worst Case 2 alone does not support interaction effects. Plus, the fused hybrid model is more robust in the face of attribute prohibitions than CBC.

In summary, we are more enthusiastic about CBC than Best-Worst Case 2 if the primary goal is to predict choice behavior. Best-Worst Case 2 leads to different utility scores, involves declaring which levels are most and least preferred rather than deriving the preferences, and does slightly worse in predicting holdout CBC tasks compared to CBC. Despite the drawbacks, fusing Best-Worst Case 2 with conjoint judgments (Best-Worst Case 3) integrates across multiple choice contexts, justifies a compositional rule, makes the results more robust (if prohibitions are involved), and places all the utility scores on the same scale. This latter benefit means the results can reveal additional strategic insights and can be more easily understood by general audiences. These benefits may outweigh the slight degradation in predictive accuracy specifically for CBC-looking contexts, leading researchers to prefer the hybrid approach for many research situations and clients. Using the hybrid approach need not involve any compromise. Researchers could use the Priors + Best-Worst Case 3 fusion to predict choices and the fusion of all three sections to place the utilities on the common scale for easier/strategic interpretation and market segmentation.

Appendix A: Designing and Coding a Sample Hybrid Survey

We demonstrate how to design a hybrid choice exercise as described in this paper and to encode the information for logit-based utility estimation (aggregate logit, latent class, or HB).

Consider a small study with three attributes, each with three levels. Further assume the first attribute is brand and the remaining two attributes are ordered attributes with a priori utility order (best to worst).

Section 1—Priors: The first section of the hybrid choice survey asks respondents to rate the three brands on, for illustration, a 4-point scale:

	1-Poor	2-OK	3-Excellent	No Opinion
Brand A	O	O	O	O
Brand B	O	O	O	O
Brand C	O	O	O	O

Section 2—Design for Best-Worst for Levels within Profiles (Best-Worst Case 2):
Using CBC's Complete Enumeration design methodology, we could select six profiles (six tasks each with one concept; or 2 task each with 3 concepts) to show in this section, so each item appears exactly twice. The six profiles in version 1 of the experimental design might be composed of the following attribute levels:

	Att1	Att2	Att3
Profile 1:	1	1	2
Profile 2:	3	2	3
Profile 3:	2	3	1
Profile 4:	1	3	3
Profile 5:	3	1	1
Profile 6:	2	2	2

Each level appears twice, and each level appears with as many other levels of other attributes as we're able to do with such few tasks. This achieves perfect level balance and a relatively high degree of connectivity. Although there is no direct connectivity among the levels within a given attribute, indirect connectivity is established—then additional direct within-attribute comparisons are provided by the Priors (Section 1) and the CBC section (Section 3).

The first task for this section looks like the following:

Which features in this product make you most and least want to purchase it? In other words, which are the best and worst aspects of this product?

	Best aspect	Worst aspect
Brand A	O	O
Speed 1	O	O
Price 2	O	O

Section 3—Design for Best-Worst CBC (Best-Worst Case 3):
For illustration, we select 5 tasks, each with three concepts. We use the Balanced Overlap design approach (which attempts to optimize level balance and orthogonality, while permitting a modest degree of level overlap within tasks). For version 1, the 5 tasks might look like the following:

Task	Concept	Brand	Speed	Price
1	1	1	1	2
1	2	3	2	3
1	3	2	3	2
2	1	1	2	3
2	2	1	3	1
2	3	3	1	1
3	1	2	1	3
3	2	3	3	2
3	3	2	2	1
4	1	2	3	2
4	2	3	2	1
4	3	1	1	3
5	1	3	3	3
5	2	2	1	1
5	3	2	2	3

The Balanced Overlap designer adds level overlap within tasks, to support robust interactions estimation. However, it does sacrifice some degree of level balance to accomplish this. Across respondents, however, the level balance is nearly perfect.

The first of the five choice tasks would look something like the following:

Considering just these three products, which is the best and which is the worst?

	Brand A	Brand C	Brand B
	Speed 1	Speed 2	Speed 3
	Price 2	Price 3	Price 2
Best	O	O	O
Worst	O	O	O

Given what you know about the market, how likely are you to buy the product you selected as best above?

O Definitely would buy
O Probably would buy
O Might or might not buy
O Probably would not buy
O Definitely would not buy

The dual-response None format is used, and the researcher must choose a cut-off point on the 5-point Likert scale that indicates the None threshold, such as top box or top two box.

Summary:

The respondent makes 30 clicks to complete the entire survey:

Priors Section: 3 clicks
Best-Worst within Profiles Section: 12 clicks
Best-Worst CBC question: 15 clicks

Description of Data Coding:

Priors Section (Section 1):

The respondent provides ratings for levels of brand on a 4-point scale. The rank-orders of the remaining two attributes (speed and price) are already known and do not need to be asked.

Assume the respondent rates the brands as:

Brand A:	2
Brand B:	3
Brand C:	3

From these relative ratings we can infer the following pairwise choices:

Brand A < Brand B
Brand A < Brand C

For the Speed and Price attributes, we can infer the following partial chain of inequalities:

Speed 1 > Speed 2
Speed 2 > Speed 3
Price 1 > Price 2
Price 2 > Price 3

We can therefore encode the choice tasks for the Priors section with six pairwise choice tasks.

We will identify the model by arbitrarily choosing Brand A to be the reference point within the dummy coding. Brand A is set to a utility of 0 in the estimation and all other levels, including None, are estimated with respect to it.

Task	Concept	BrB	BrC	Spd1	Spd2	Spd3	Pr1	Pr2	Pr3	None	Choice
1	1	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	0	0	0	0	0	0	1
2	1	0	0	0	0	0	0	0	0	0	0
2	2	0	1	0	0	0	0	0	0	0	1
3	1	0	0	1	0	0	0	0	0	0	1
3	2	0	0	0	1	0	0	0	0	0	0

4	1	0	0	0	1	0	0	0	0	0	1
4	2	0	0	0	0	1	0	0	0	0	0
5	1	0	0	0	0	0	1	0	0	0	1
5	2	0	0	0	0	0	0	1	0	0	0
6	1	0	0	0	0	0	0	1	0	0	1
6	2	0	0	0	0	0	0	0	1	0	0

Best-Worst Choices of Levels within Profiles (Section 2):

Let's assume the respondent indicated that within the first profile of this section [Brand A, Speed 1, Price 2] that Speed 1 is the best aspect and Brand A is the worst aspect of this profile.

We encode the information as two tasks, a best task (Task 1b) and a worst task (Task 1w). Within the worst task, the design matrix is multiplied by -1:

Task	Concept	BrB	BrC	Spd1	Spd2	Spd3	Pr1	Pr2	Pr3	None	Choice
1b	1	0	0	0	0	0	0	0	0	0	0
1b	2	0	0	1	0	0	0	0	0	0	1
1b	3	0	0	0	0	0	0	1	0	0	0
1w	1	0	0	0	0	0	0	0	0	0	1
1w	2	0	0	-1	0	0	0	0	0	0	0
1w	3	0	0	0	0	0	0	-1	0	0	0

Best-Worst tasks 2 through 5 are coded in the same way, leading to a total of 10 choice tasks coded for this section within the hybrid survey.

Best-Worst CBC Section (Section 3):

This final section adds more information about the relative preferences for levels as well as providing the conjoint measurement section of the hybrid (lending formal support to the compositional rule).

In Task 1 of this section, the respondent sees the following question:

> Considering just these three products, which is the
> best and which is the worst?

	Brand A Speed 1 Price 2	Brand C Speed 2 Price 3	Brand B Speed 3 Price 2
Best	O	O	O
Worst	O	O	O

> Given what you know about the market,
> how likely are you to buy the product you selected
> as best above?
>
> O Definitely would buy
> O Probably would buy

O Might or might not buy
O Probably would not buy
O Definitely would not buy

The treatment of dual-response None data is described in the CBC/HB documentation. To summarize, if the respondent indicates he "would buy" the concept he selected as best, then we include the None alternative as an available concept within the "best" task, but indicate that it isn't selected. If the respondent indicates that he "would not buy" the concept he selected as best, then we need to code an additional task. The None concept only enters into the additional task, with it being selected instead of the three product concepts shown within the Task.

Let's assume the respondent picks the first concept as best, the third concept as worst, and indicates that he probably would buy the concept he selected as best. Furthermore assume that the analyst has indicated that top-two box indicates a "buy" and bottom three boxes indicate "no buy" or the None choice.

We illustrate the coding for the task above:

Task	Concept	BrB	BrC	Spd1	Spd2	Spd3	Pr1	Pr2	Pr3	None	Choice
1b	1	0	0	1	0	0	0	1	0	0	1
1b	2	0	1	0	1	0	0	0	1	0	0
1b	3	1	0	0	0	1	0	1	0	0	0
1b	4	0	0	0	0	0	0	0	0	1	0
1w	1	0	0	-1	0	0	0	-1	0	0	0
1w	2	0	-1	0	-1	0	0	0	-1	0	0
1w	3	-1	0	0	0	-1	0	-1	0	0	1

If the respondent had instead said that he probably would not buy the product he chose as best, we would encode the information in three tasks, where task 1n indicates that the None is preferred to all three concepts:

Task	Concept	BrB	BrC	Spd1	Spd2	Spd3	Pr1	Pr2	Pr3	None	Choice
1b	1	0	0	1	0	0	0	1	0	0	1
1b	2	0	1	0	1	0	0	0	1	0	0
1b	3	1	0	0	0	1	0	1	0	0	0
1w	1	0	0	-1	0	0	0	-1	0	0	0
1w	2	0	-1	0	-1	0	0	0	-1	0	0
1w	3	-1	0	0	0	-1	0	-1	0	0	1
1n	1	0	0	1	0	0	0	1	0	0	0
1n	2	0	1	0	1	0	0	0	1	0	0
1n	3	1	0	0	0	1	0	1	0	0	0
1n	4	0	0	0	0	0	0	0	0	1	1

All Three Sections Coded Together for Parameter Estimation:

We code all three sections of the hybrid choice task together within a single data file for parameter estimation. Formally, there is something incorrect about doing this, since the response error (and therefore the scale factor) will be different per section. There are at least three things that could be done about this:

1. Do nothing. The final parameters will be a compromise among the different underlying parameters for the three different sections as well as the magnitude of the parameters in terms of the scale factors. The compromise will provide a high likelihood fit to the data, but not as good of fit if unique scale parameters were also estimated for two of the sections relative to a third section (typically the Best-Worst CBC section) fixed at a scale of 1.0.
2. Simultaneously estimate the parameters and the scale factors of two sections relative to the third section, where the third section (typically the Best-Worst CBC Section) is fixed at a scale of 1.0. Such a solution is provided by Otter (Otter 2007).
3. Proceed as with step 1, but do an aggregate (one scale parameter for the entire sample) post hoc tuning of the scale factor to best fit the scale of the Best-Worst CBC section (to maximize LL).

Interaction Effects:

One of the benefits of the hybrid method we describe is the ability to include interaction effects within the parameter estimation. The interaction effects are only estimable from the Best-Worst CBC section of the interview, so the columns encoding the interaction effects within the design matrix remain as 0s for the rows of the other two sections.

We would prefer that the main effects remain interpretable independent of the interaction effects, so we estimate the interaction effects using effects-coding (rather than dummy coding) so that the interaction effects are zero-centered.

Consider that we wish to include the interaction effects between brand and price within the estimation. For Best-Worst CBC task 1, the three concepts shown are:

Brand A	Brand C	Brand B
Speed 1	Speed 2	Speed 3
Price 2	Price 3	Price 2

If we were coding the main effects parameters as effects coding in a standard CBC questionnaire, we would code just the brands and prices as:

Task	Concept	BrB	BrC	Pr2	Pr3
1	1	-1	-1	1	0
1	2	0	1	0	1
1	3	1	0	1	0

And, the interactions between brand and price would additionally be encoded (adding new columns to the design matrix) by cross-multiplying the columns between the two attributes:

Task	Concept	BrB X Pr2	BrB x Pr3	BrC x Pr2	BrC x Pr3
1	1	-1	0	-1	0
1	2	0	0	0	1
1	3	1	0	0	0

These four columns are the additional encoded variables to capture interaction effects between brand and price that need to be added to the Best-Worst CBC design matrix. Because the data matrix must contain the same number of columns for the encoding of all three sections (since the parameter estimation occurs within a single pooled model), these same four columns exist for the other three sections of the survey, and contain 0s as entries for the rows.

After estimation, the interaction effects between brand and price may be "expanded" to explicitly show all nine zero-centered interaction effects. This is illustrated in the table below. In standard CBC practice, the interaction between two 3-level attributes is coded as (3-1)(3-1)=4 columns. In the table below, parameters A, B, C, and D are explicitly estimated as utility scores. A is the interaction effect between BrB and Pr1. B is the interaction effect between BrC and Pr1, etc. The remaining five interaction effects may be inferred and expanded using the simple formulas below.

	BrA	BrB	BrC
Pr1	-(A+B)	A	B
Pr2	-(C+D)	C	D
Pr3	-(A+B+ C+D)	-(A+C)	-(B+D)

To summarize, if applying interaction effects within the utility estimation for our common scale hybrid choice method, all main-effect parameters are placed on a common scale, where one of the levels has been chosen as the zero point. Interaction terms between attributes taken two-at-a-time may be estimated using effects-coding, such that the interaction effects are zero-centered and do not bother the interpretability of the main effects parameters.

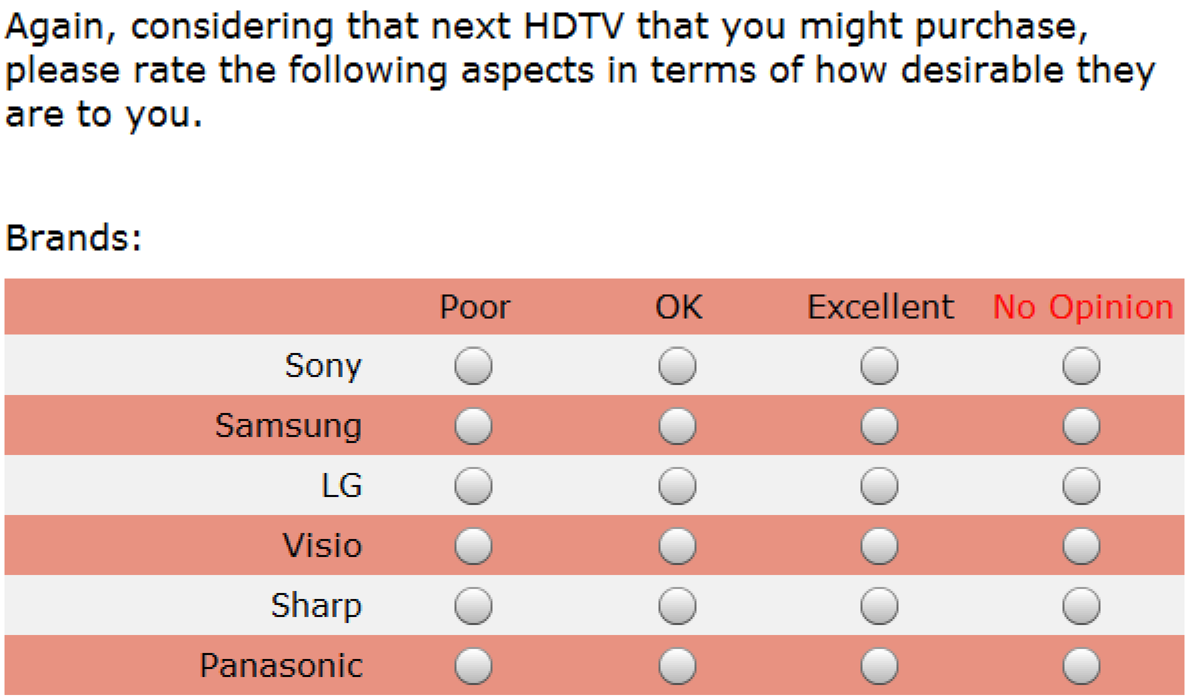
The interaction effects are mainly of value within the market simulator and do not need to be presented to less sophisticated audiences that may be satisfied by a reporting of the main effects on the common scale.

Presenting the Results as Positive Values on a Common Scale:

It may make it even easier to present the results if the main effect scores are rescaled to a ratio scale that has a theoretical low mark of 0 and where the scores sum to 100. This may be done by first zero-centering the utilities and then transforming each by $e^{Ui}/(1+e^{Ui})$. Finally, for convenience, the scores may be rescaled by making them sum to 100.

Appendix B: Screen Shots from HDTV Study

A portion of the Priors rating grid (for just the Brand attribute):

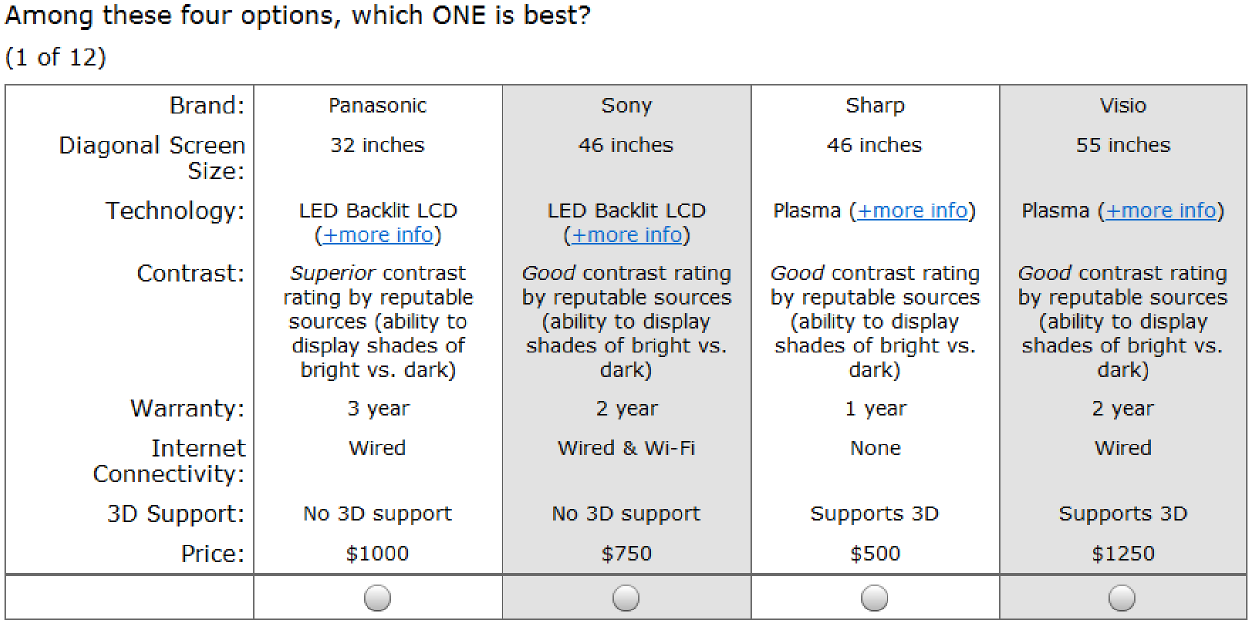


Again, considering that next HDTV that you might purchase, please rate the following aspects in terms of how desirable they are to you.

Brands:

	Poor	OK	Excellent	No Opinion
Sony	○	○	○	○
Samsung	○	○	○	○
LG	○	○	○	○
Visio	○	○	○	○
Sharp	○	○	○	○
Panasonic	○	○	○	○

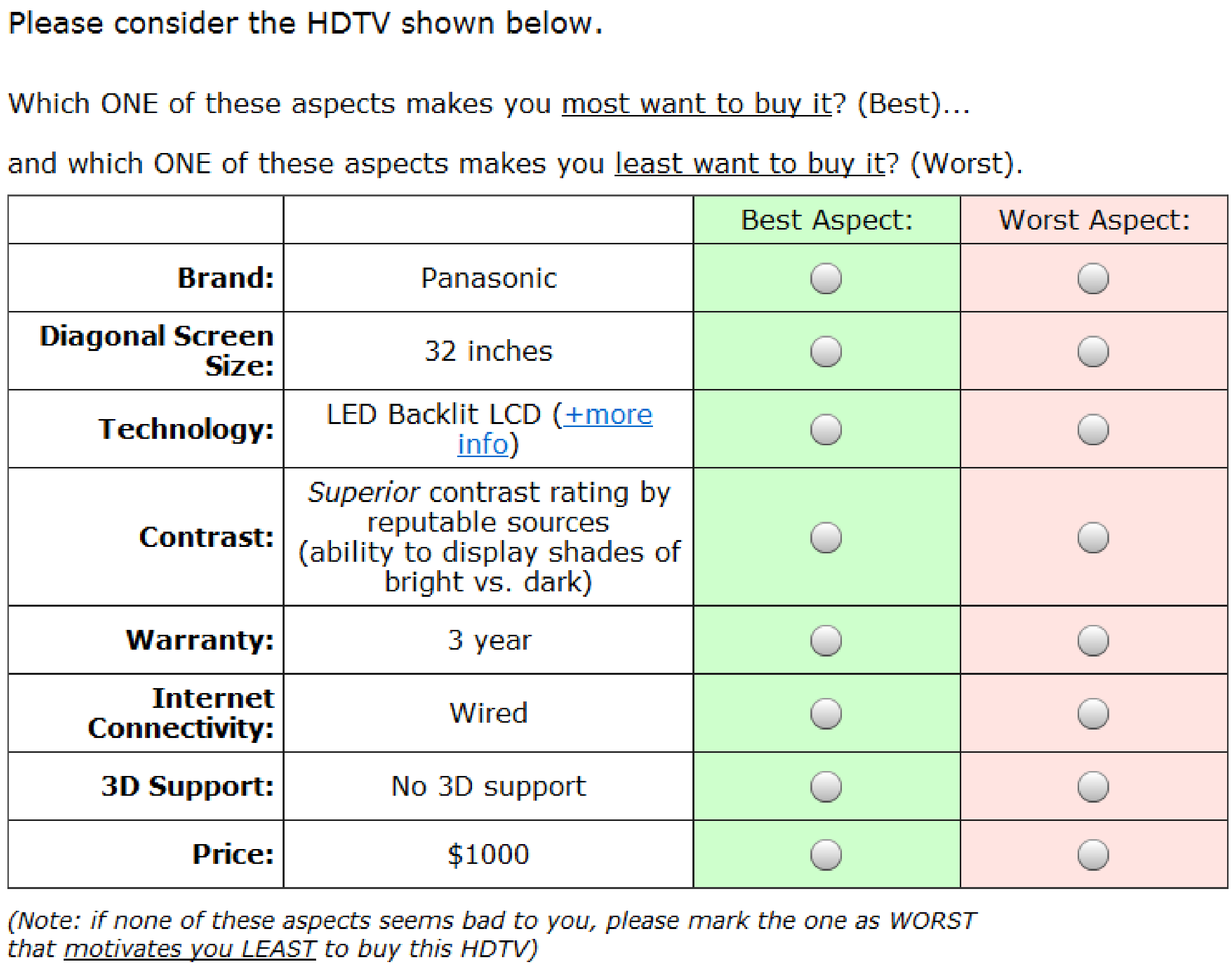
CBC Tasks (used for Cell 1 Holdouts and Cell 2 Utility Estimation):



Among these four options, which ONE is best?
(1 of 12)

Brand:	Panasonic	Sony	Sharp	Visio
Diagonal Screen Size:	32 inches	46 inches	46 inches	55 inches
Technology:	LED Backlit LCD (+more info)	LED Backlit LCD (+more info)	Plasma (+more info)	Plasma (+more info)
Contrast:	*Superior* contrast rating by reputable sources (ability to display shades of bright vs. dark)	*Good* contrast rating by reputable sources (ability to display shades of bright vs. dark)	*Good* contrast rating by reputable sources (ability to display shades of bright vs. dark)	*Good* contrast rating by reputable sources (ability to display shades of bright vs. dark)
Warranty:	3 year	2 year	1 year	2 year
Internet Connectivity:	Wired	Wired & Wi-Fi	None	Wired
3D Support:	No 3D support	No 3D support	Supports 3D	Supports 3D
Price:	$1000	$750	$500	$1250
	○	○	○	○

Full-Profile MaxDiff Case 2 Tasks:



Please consider the HDTV shown below.

Which ONE of these aspects makes you most want to buy it? (Best)...

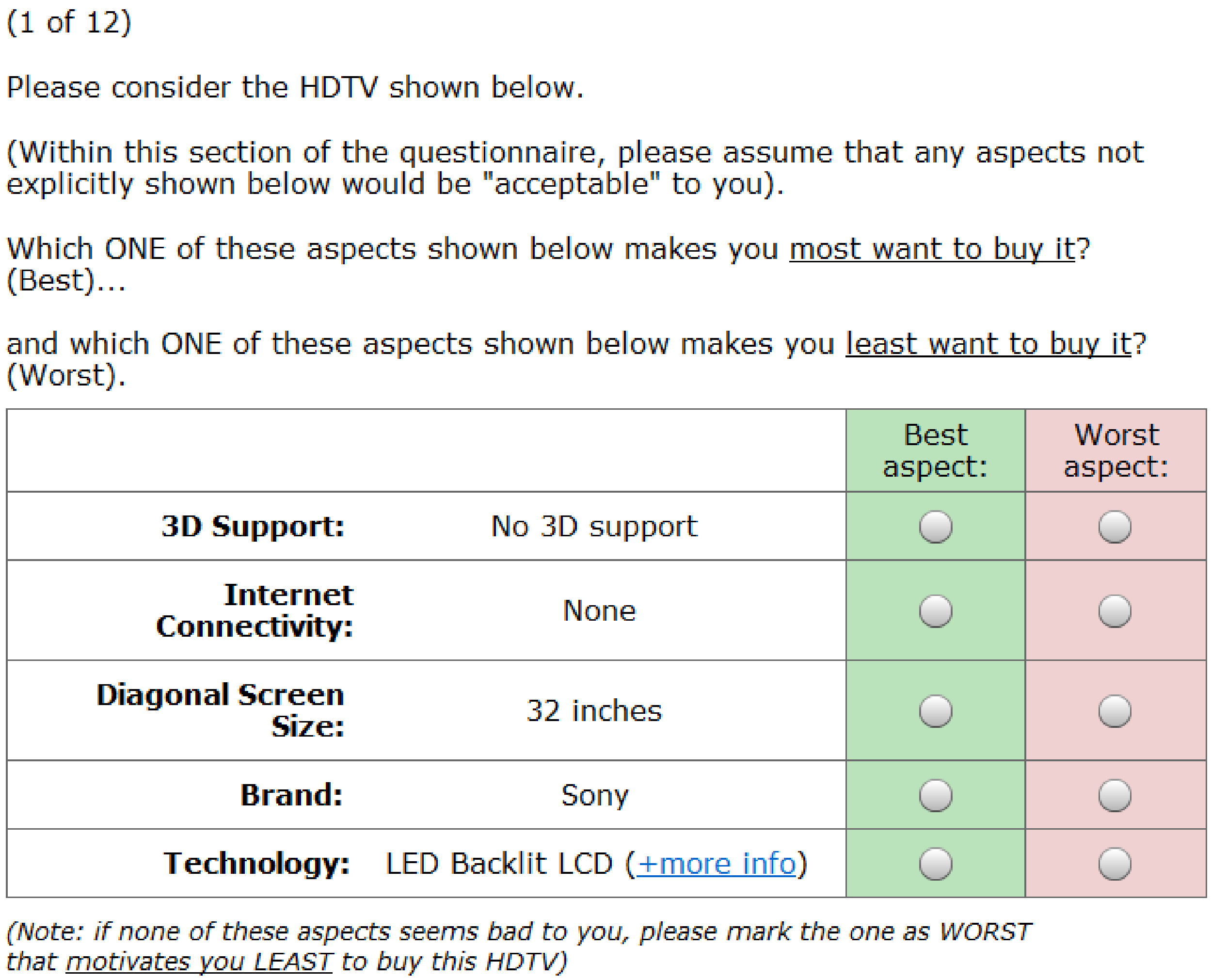
and which ONE of these aspects makes you least want to buy it? (Worst).

		Best Aspect:	Worst Aspect:
Brand:	Panasonic	○	○
Diagonal Screen Size:	32 inches	○	○
Technology:	LED Backlit LCD (+more info)	○	○
Contrast:	*Superior* contrast rating by reputable sources (ability to display shades of bright vs. dark)	○	○
Warranty:	3 year	○	○
Internet Connectivity:	Wired	○	○
3D Support:	No 3D support	○	○
Price:	$1000	○	○

(Note: if none of these aspects seems bad to you, please mark the one as WORST that motivates you LEAST to buy this HDTV)

(After a few questions, we shortened the explanation text and removed the note from the bottom of the task. After respondents pick best and worst levels, we remove those two attributes and show a profile involving the remaining six attributes. Respondents are then asked to select best and worst levels among the remaining six attributes.)

Partial-Profile MaxDiff Case 2 task:



(1 of 12)

Please consider the HDTV shown below.

(Within this section of the questionnaire, please assume that any aspects not explicitly shown below would be "acceptable" to you).

Which ONE of these aspects shown below makes you most want to buy it? (Best)...

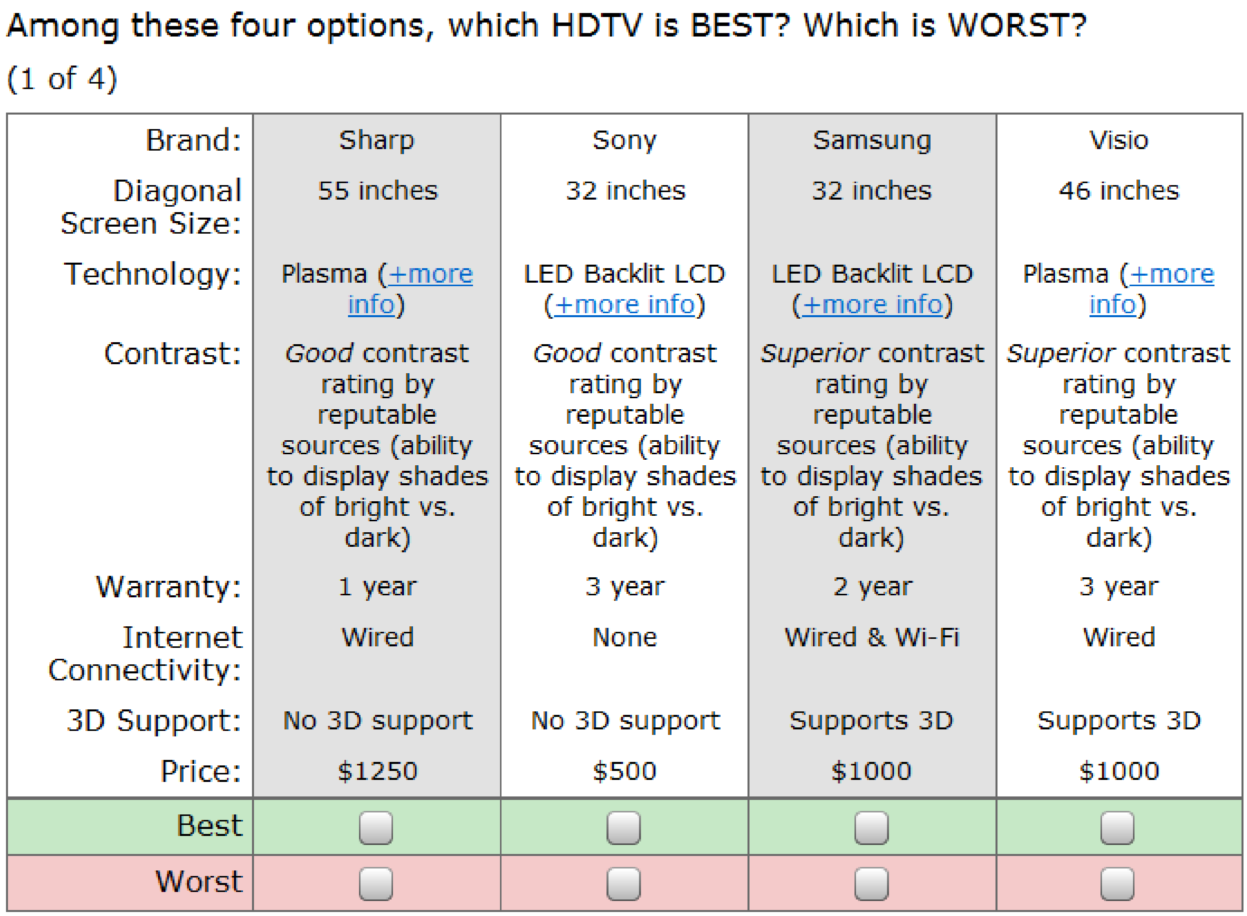
and which ONE of these aspects shown below makes you least want to buy it? (Worst).

		Best aspect:	Worst aspect:
3D Support:	No 3D support	○	○
Internet Connectivity:	None	○	○
Diagonal Screen Size:	32 inches	○	○
Brand:	Sony	○	○
Technology:	LED Backlit LCD (+more info)	○	○

(Note: if none of these aspects seems bad to you, please mark the one as WORST that motivates you LEAST to buy this HDTV)

(After a few questions, we shortened the explanation text and removed the note from the bottom of the task.)

Best-Worst Case 3 Task:



Among these four options, which HDTV is BEST? Which is WORST?

(1 of 4)

Brand:	Sharp	Sony	Samsung	Visio
Diagonal Screen Size:	55 inches	32 inches	32 inches	46 inches
Technology:	Plasma (+more info)	LED Backlit LCD (+more info)	LED Backlit LCD (+more info)	Plasma (+more info)
Contrast:	*Good* contrast rating by reputable sources (ability to display shades of bright vs. dark)	*Good* contrast rating by reputable sources (ability to display shades of bright vs. dark)	*Superior* contrast rating by reputable sources (ability to display shades of bright vs. dark)	*Superior* contrast rating by reputable sources (ability to display shades of bright vs. dark)
Warranty:	1 year	3 year	2 year	3 year
Internet Connectivity:	Wired	None	Wired & Wi-Fi	Wired
3D Support:	No 3D support	No 3D support	Supports 3D	Supports 3D
Price:	$1250	$500	$1000	$1000
Best	☐	☐	☐	☐
Worst	☐	☐	☐	☐

References

Chrzan K. and M. Skrapits (1996), "Best-Worst Conjoint Analysis: An Empirical Comparison with a Full-Profile Choice-Based Conjoint Experiment," paper presented at the INFORMS Marketing Science Conference, Gainesville, FL.

Chrzan, Keith, John Zepp, and Joseph White (2012), "The Success of Choice-Based Conjoint Designs among Respondents Making Lexicographic Choices," 2010 Sawtooth Software Conference Proceedings, Sequim, WA, pp 19-35.

Elder, Andrew and Terry Pan (2009), "Survey Quality and MaxDiff: An Assessment of Who Fails, and Why," 2009 Sawtooth Software Conference Proceedings, Sequim, WA, 55-82.

Finn, A. and J. J. Louviere (1992), "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," Journal of Public Policy and Marketing, 11, 1, 12-25.

Flynn, Terry N., Tim J. Peters, and Joanna Coast, "Quantifying Response Shift or Adaptation Effects in Quality of Life by Synthesizing Best-Worst Scaling and Discrete Choice Data," The Journal of Choice Modelling, 6 (2013) 34-43.

Grimshaw, Scott D., Bruce J. Collings, Wayne A. Larsen, and Carolyn R. Hurt (2001), "Eliciting Factor Importance in a Designed Experiment," Technometrics, May 2001, Vol. 43, No. 2.

Hoogerbrugge, Marco and Kees van der Wagt (2007), "How Many Choice Tasks Should We Ask?" Sawtooth Software Conference Proceedings, pp 97-110, Sequim, WA.

Kurz, Peter and Stefan Binner (2012), "'The Individual Choice Task Threshold' Need for Variable Number of Choice Tasks," Sawtooth Software Conference Proceedings, Orem, UT, pp 111-128.

Lattery, Kevin (2009), "Coupling Stated Preferences with Conjoint Tasks to Better Estimate Individual-Level Utilities," Sawtooth Software Conference, Sequim, WA.

Lattery, Kevin and Bryan Orme (2012), "Can We Improve CBC Questionnaires with Strategically-Placed Level Overlap and Appropriate 'Screening Rule' Questions?" 2012 Sawtooth Software Conference Proceedings, Orem, UT, pp 313-332.

Lenk, Peter and Bryan Orme (2009), "The Value of Informative Priors in Bayesian Inference with Sparse Data," with Bryan Orme, Journal of Marketing Research, 46, 6, 832-845.

Louviere, J. J. (1991), "Best-Worst Scaling: A Model for the Largest Difference Judgments," Working Paper, University of Alberta.

Louviere, J. J. (1993), "The Best-Worst or Maximum Difference Measurement Model: Applications to Behavioral Research in Marketing," The American Marketing Association's 1993 Behavioral Research Conference, Phoenix, Arizona.

Louviere, J. J., and G. G. Woodworth (1983), "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," Journal of Marketing Research 20, 350-367.

Louviere, J. J., Joffre Swait, and Donald Anderson (1995), “Best/Worst Conjoint: A New Preference Elicitation Method to Simultaneously Identify Overall Attribute Importance and Attribute Level Partworths,” Unpublished working paper, University of Sydney.

Marley, A., Terry Flynn, Jordan Louviere (2008), “Probabilistic Models of Set-Dependent and Attribute-Level Best-Worst Choice,” Journal of Mathematical Psychology 52, 281-296.

Marshall, Don, Siu-Shing Chan, and Joseph Curry (2010), “A Head-to-Head Comparison of the Traditional (Top-Down) Approach to Choice Modeling with a Proposed Bottom-Up Approach,” 2010 Sawtooth Software Conference Proceedings, Sequim, WA, pp 309-320.

Orme, Bryan (2005), “Accuracy of HB Estimation in MaxDiff Experiments,” Sawtooth Software Research Paper Series, available at www.sawtoothsoftware.com.

Orme, Bryan (2010), “Comment on Marshall et al. and Wirth,” 2010 Sawtooth Software Conference Proceedings, Sequim, WA, pp 357-361.

Otter, Thomas (2007), “HB-Analysis for Multi-Format Adaptive CBC,” Sawtooth Software Conference Proceedings, Sequim, WA.

Sagal, Christopher and Ely Dahan (2012), “The Voice of the Patient,” Sawtooth Software Conference Proceedings, Orem, Utah.

Sawtooth Software (1993), CBC System for Choice-Based Conjoint.

Sawtooth Software (1998), “On Interaction Effects and CBC Designs,” *Sawtooth Solutions*, Fall 1998.

Sawtooth Software (2004), MaxDiff System for Maximum Difference Scaling.

Sawtooth Software (2008), ACBC System for Adaptive Choice-Based Conjoint.

Sawtooth Software (2012), MBC System for Multi-Check Choice Experiments.

Srinivasan and Park (1997), “Surprising Robustness of the Self-Explicated Approach to Customer Preference Structure Measurement,” Journal of Marketing Research, May 1997, 286-291.

Tang, Jane and Andrew Grenville (2010), “How Many Questions Should You Ask in CBC Studies?—Revisited Again,” Sawtooth Software Conference Proceedings, Sequim, WA.

Uldry, Pierre; Valerie Severin and Chris Diener. “Using A Dual Response Framework in Choice Modelling,” 2002 AMA Advanced Research Techniques Forum, Vail, Colorado.

Wirth, Ralph (2010), “HB-CBC, HB-Best-Worst-CBC or No HB at All?” 2010 Sawtooth Software Conference Proceedings, Sequim, WA, pp 321-356.