Sawtooth Software

RESEARCH PAPER SERIES

Assessing the Validity of Conjoint Analysis – Continued

Bryan K. Orme,
Sawtooth Software, Inc.,
Mark I. Alpert,
The University of Texas at Austin
and
Ethan Christensen,
The University of Texas at Arlington,
1997

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

ASSESSING THE VALIDITY OF CONJOINT ANALYSIS—CONTINUED

Bryan K. Orme
Sawtooth Software, Inc.

Mark I. Alpert
The University of Texas at Austin

Ethan Christensen
The University of Texas at Arlington

INTRODUCTION

Despite over 20 years of conjoint research and hundreds of methodological papers, very little has been published in the way of formal tests of whether conjoint *really* works in predicting significant *real-world* actions. As ancillary cases, there are interesting questions of whether one method works better than another, and under what circumstances each method should be preferred.

Most of the debate to this point has focused on calibration of utilities. Our research focuses on the other side of the equation: the validity measurement. In most validity studies, researchers have begged off the measurement of real validity by settling for attempts to predict holdout concepts administered in the same interview. Because the holdout concept is usually so similar (even identical) to the conjoint exercise, most validity studies really only measure internal consistency. When viewed with any perspective at all, calling such exercises validity studies seems a presumptuous stretch. Further, in the typical conjoint validity study, as much as 95% of the effort goes into measuring respondent utilities, and as little as 5% goes into measuring what it is we want to predict. It seems as though validity studies should invest much more in measurement of that which is to be predicted.

There are some widely-recognized shortcomings of conjoint methods. For example, it is thought that respondents sometimes use simplification strategies to answer difficult full-profile tasks. Respondents may consider only the few most important attributes, which would result in exaggerated differences in importance between the most and least important factors. And it is also thought that ACA sometimes errs in forcing individuals to pay attention to every attribute, whether important or not, which would result in ACA's importances being "too flat." Of course, lacking a proper validity study based on real-world purchase observation, both of these claims remain only conjecture.

The title for our research is taken from a paper entitled "Assessing the Validity of Conjoint Analysis" presented by Rich Johnson in the 1989 Sawtooth Software Conference Proceedings. We refer to important points from that paper, and then report an original pilot study, which attempts to overcome some of the weaknesses of traditional validation research. This pilot study featured an intensive

---

Dr. Mark Alpert holds the Foley's Centennial Professorship in Marketing at The University of Texas at Austin. Ethan Christensen is a doctoral student at The University of Texas at Arlington. We thank Rich Johnson and Joel Huber for their insightful comments and direction. The authors accept responsibility for any errors.

holdout exercise, which may better reflect real world purchase behavior than traditional holdouts asked during the course of conjoint surveys.

Our main emphasis is on the principles of design for conjoint validity studies. We also compare the results from full-profile, ACA and choice-based conjoint. Our results are not powerful enough to reach strong conclusions about methods, but we think we illustrate a way for strengthening traditional validity studies. Finally, we note the limitations of our pilot study and suggest directions for additional research.

DESIGN CONSIDERATIONS FOR HOLDOUT TASKS

In conjoint validation studies, holdout tasks are not used in the estimation of part-worths. They are presumed to represent how the respondent would choose in the real world. Researchers measure validity by comparing how well conjoint utilities predict choices from the holdout tasks.

Ideally, the actual purchase event should be the criterion measurement. Lacking real purchase data, guidelines for constructing experimental holdout tasks include:

- At least one of the holdout tasks should be repeated to assess the reliability of holdout judgements. This allows the researcher to determine the proportion of error in prediction due to errors in the part-worths versus errors in response to holdout judgements themselves. It also provides a way to adjust hit rates when comparing results from independent samples of not necessarily the same response reliability.

- The validation measurement should closely mimic the stimulus presentation and depth of processing of the real world purchase event.

- Attribute order effects should be controlled. The holdout task should present the attributes in a different order than full-profile calibration tasks.

HOW REALISTIC ARE TRADITIONAL HOLDOUT TASKS?

The majority of conjoint validity tests have used full-profile evaluations as holdout tasks. Hit rates for correctly predicted choices (from choice sets of pairs, triples, etc.) are a popular measure, as well as correlation or MSE for ratings-based holdouts. It has been argued that full-profile holdouts best represent how products are viewed and evaluated in the real world. We think this is reasonable, but we question whether buyers process and evaluate full-profiles in the real world the same way they do during the context of a survey.

Particularly for high-involvement purchases, respondents exert more effort making real-world decisions than while making judgments in conjoint surveys. Once warmed up to the task, respondents can take as little as 12 seconds on average to make choices in full-profile choice questionnaires (Johnson and Orme, 1996). Huber observes: "Purchases of laptops are generally not made in anything like 30 seconds; people spend significant time discussing a wide range of features" (Huber *et al*. 1992).

Full-profile interviews involving many attributes may encourage respondents to adopt simplification heuristics. By focusing on just a subset of the attributes, respondents can more easily complete long

and monotonous conjoint interviews. Simplification strategies can lead to more extreme attribute importances, with relatively little weight given to the factors the respondent has chosen to ignore. If simplification heuristics are indeed being used in full-profile judgements, it would lead to some critical questions:

- Do respondents focus on just a subset of key attributes in the real world, when many attributes are involved and real dollars are on the line?
- Are hastily-answered full-profile holdout concepts realistic criteria for measuring conjoint validity?
- If not, what criteria should we use for validity comparisons?

These questions form the crux of our research. Again, the ideal validation study would attempt to predict actual purchase behavior. In the absence of real-world judgements, other steps might be taken to improve the quality of the holdout task. For our study, we designed a "Super Holdout Task" (described below) to address in part the shortcomings of tradition holdouts and better simulate real world behavior.

We hypothesize that especially with significant decisions, individuals may broaden their range of attention to product features, perhaps resulting in flatter importances than are captured with traditional full-profile holdout concepts. With this hypothesis in mind, we turn to details of our study design.

STUDY DESIGN

MBA's from The University of Texas at Austin, The University of Texas at Arlington and the University of Washington were employed as respondents. The subject of the study was personal computers for a hypothetical new computer lab at the respondents' respective business schools.

Nine attributes were studied: brand, warranty, microprocessor, number of lab assistants, ergonomic keyboard/mouse, hard drive, RAM, modem/Internet access, and price. All attributes had either two or three levels described in succinct phrases. (See Appendix A for a full listing of attribute levels).

There were two main components of the study, administered in two separate sessions:

1) Computer-Administered Survey. Respondents received a packet that included a survey disk programmed using Ci3 along with 22 full-profile conjoint cards printed on card-stock paper. The order of tasks in the survey was:

a) Demographic questions. (Experience with and familiarity with personal computers/ past purchase influence for PCs.)

b) Full-profile card-sort/ACA. (Each respondent received both, in rotated order). Sawtooth Software's CVA system was used to design and analyze the full-profile data. Twenty-two hard-copy cards were sorted into four piles based on preference, and then rated using a 100-pt scale. To control for attribute order bias, two versions of the cards were printed. (See Appendix B for details of the full-profile design).

ACA v4.0 was used with default settings.

c) Five full-profile holdout choice tasks with three product concepts each. These tasks were constructed randomly using Ci3. Respondents indicated first and second choices. The second and fifth tasks were identical (with rotated concepts) to measure test-retest reliability. Brand was always the first attribute, and Price the last. The interior seven attributes were randomly rotated across respondents.

2) Super Holdout Task. After completing the disk-based survey, respondents participated in a 10-minute in-class exercise. Students divided into committees of three to evaluate just one choice task with four PC configurations described in full-profile. These tasks were constructed randomly, varying across committees. Attribute order was randomized for the interior seven attributes.

Each committee was instructed to reach consensus regarding the best, second, third and worst PC configuration for the new computer lab. After the group evaluations were recorded, respondents were asked to record their personal evaluations—but they did not know beforehand that we would ask for their personal opinions.

We expected that the Super Holdout Task might better reflect real-world behavior than traditional validation tasks, particularly for high-involvement categories. Since more is at stake with high-dollar purchases, buyers spend a great deal of time weighing the pros and cons of available alternatives. Buyers also seek additional information by consulting with others. Also, for business-to-business markets (or even for households), purchase decisions are frequently decided by some sort of committee after some debate.

Another unique aspect of this study is the use of randomized holdout tasks. For many-attribute designs (such as ours) the randomized approach generally achieves a fair degree of utility balance, which is a desirable condition for testing predictive validity. Consistently dominated choice tasks would be less useful, since predicting choices for dominated tasks is trivial. Perhaps most useful is that randomized designs permit group-level utility estimation for the holdout judgments. We are able to compare utilities and importances from four sources: full-profile ratings, ACA, standard holdout choices, and the Super Holdout Task. We expect to assess whether part-worths differ between traditional holdouts and the Super Holdout Task.

Based on previous research (Huber 1992, Pinnell 1994), we expected that attribute importances from full-profile and full-profile choice would be more extreme than ACA importances. We also hypothesized that importances derived from the Super Holdout Task would be less extreme than the holdout choice exercise that was part of the computer-administered interview, reflecting greater depth of processing.

TIMING DATA AND RESPONDENT PROFILE

A total of 80 completed surveys were analyzed. Median interview time was 27 minutes for the disk survey, including 13 minutes for the full-profile exercise (time to sort and write scores on the cards), and 8 minutes for ACA. ACA took significantly less time to complete than full-profile, with a t-value for the mean difference in interview time of 6.9.

There were five standard holdout choice tasks during the computer-administered portion of the

survey. These took 48, 35, 32, 28 and 26 seconds to complete. Although we don't have timing data for the Super Holdout Task, it took respondents about 10 minutes to complete.

Seventy-four percent of the respondents had been the main decision-maker for purchasing a PC before.

CALCULATING HOLDOUT HIT RATES

Conjoint validity is usually assessed by observing how well part-worths can predict holdout evaluations. First choice hits are the most common measure. For choice tasks including judgments beyond first choice, we may evaluate hit rates for an expanded set of implied comparisons. For the choice-based holdout tasks in our study, we asked for a full ranking of alternatives. The choice-based tasks in the disk survey involved three product concepts. Assuming the preference order was a, b, c, there are three implied inequalities: a>b, a>c, b>c. The Super Holdout Task presented a choice-based set with four concepts, leading to six implied inequalities (assuming preference order of a, b, c, d): a>b, a>c, a>d, b>c, b>d, c>d.

As mentioned previously, it is desirable to repeat at least one of the holdout concepts to assess test-retest reliability. This would be especially critical if, for instance, one group of respondents had completed ACA and the other group completed full-profile. In order to determine whether the conjoint method that one group received performed better than the other, we would need to adjust hit rates by the test-retest reliability for each group.

For our study, each respondent completed an identical set of calibration tasks (but with rotated task order), so we do not need to be concerned with comparing hit rates across independent samples.

Repeated holdout tasks permit us to calculate a theoretical upper limit for holdout predictability. The second and fifth choice tasks in the disk-based survey were identical (but with concepts rotated). The test-retest reliability for all implied inequalities was 90.0%. Wittink and Johnson (1992) demonstrated that the maximum expected hit rate for predicting a fallible criterion measure is equal to:

$$\pi = \frac{1 + \sqrt{(2p - 1)}}{2}$$

where π is the maximum expected hit rate and p is the agreement between independent replications of the criterion measure. Given test-retest reliability of 90.0%, the maximum possible hit rate for predicting the standard holdout choices is 94.7%. Since we did not repeat holdout tasks for the Super Holdout Task, we cannot compute its test-retest reliability. Future studies could administer a replication of the Super Holdout Task during the course of the survey to permit test-retest reliability adjustments in the case of independent samples.

TRADITIONAL HOLDOUT HIT RATES

Hit rates for full-profile and ACA are provided in the table below. Due to the small sample size, hit rates for first choices were not very stable. Hit rates for all implied inequalities included more information per respondent, and are used throughout the remainder of this paper.

OLS part-worths were calculated for full-profile. *A priori* attributes were constrained to remedy sign reversals using CVA's tieing algorithm. We used a logit transform of the 100-point purchase likelihood scale response.

Table 1

Traditional Holdout Choice Hit Rates

ACA	76.9%
Full-profile	82.4%

n=80

For the eighty respondents in the pilot study, the full-profile method does a better job predicting the standard holdout choices. The t-value for the mean difference in prediction for full-profile versus ACA for the standard holdouts is 2.85.

Why does full-profile do better than ACA for predicting the standard full-profile holdout choice tasks in our pilot study? Investigating differences between attribute importances and part-worths help answer that question.

ATTRIBUTE IMPORTANCES

We define attribute importances in the standard way, by percentaging the ranges of attribute utilities. Our study design permits us to compute attribute importances from four sources:

1) Full-profile ratings
2) ACA
3) Standard holdout choices (choice-based conjoint)
4) Super Holdout Task (choice-based conjoint)

The two choice-based sources were analyzed in the aggregate using logit, including information from all choices. In general, we suggest only using first choices within each task for utility calculation (Johnson and Orme, 1996), but the additional information was valuable for obtaining reasonable estimates given our limited sample. Also, we felt that the bias from second choices reported by Johnson and Orme would have minimal or no impact on the analysis of importances.

Respondents' enthusiasm for (or attention to) non-ordered attributes is not reflected in importances calculated from aggregate utilities when there is disagreement about which levels are preferred. Two of the nine attributes were not *a priori* ordered (brand and ergonomic keyboard/mouse) and were dropped. Relative importances for ACA, full-profile and the traditional (standard) holdout choices are shown in the table below:

Table 2

Attribute Importances

	ACA	Full-Profile	Standard Choices
Internet Access	21.5	24.6	25.7
RAM	19.8	21.6	23.8
Price	15.0	16.5	18.1
Hard Drive	14.0	14.1	12.2
Warranty	11.6	10.2	8.4
Processor	10.7	9.5	9.5
Lab Assistants	7.4	3.5	2.3
TOTAL	**100.0**	**100.0**	**100.0**
STD. DEVIATION	**4.6**	**6.8**	**7.9**

n=80

The standard deviations in the last row of the table reflect the amount of dispersion in the importances for each column. As expected, the ACA importances show the least variation from the most important to least important attributes. This confirms similar findings, which have shown ACA importances to be flatter than full-profile and choice-based results (Pinnell 1994, Huber 1992). Although the rank-order of attribute importance is identical for ACA and full-profile, the full-profile importances more closely line up with the importances derived from the traditional holdout choice exercise. This difference largely (if not principally) accounts for full-profile's edge in predicting the traditional holdouts over ACA, as will be shown below in Table 3.

The fact that full-profile closely matches the full-profile choice importances is not surprising given the similarities between the two full-profile tasks. Full-profile conjoint should have an advantage over ACA for predicting standard holdouts also shown in full-profile—especially if respondents adopt simplification heuristics.

When the ACA utilities are re-scaled at the individual level to full-profile importances, the hit rate for ACA approximates the hit rate for full-profile:

Table 3

Traditional Holdout Hit Rates
ACA Utilities Scaled to Full-profile Importances

ACA (before rescaling)	76.9%
ACA (after rescaling)	**82.6%**
Full-profile	82.4%

n=80

Scaling ACA utilities to full-profile importances significantly improves ACA's ability to predict the standard holdout choices (t=3.69).

SUPER HOLDOUT TASK RESULTS

For the Super Holdout Task, respondents were divided into groups of three individuals. Each group received a piece of paper showing four PCs described in full-profile. Over a 10-minute period, each group discussed the options and worked to consensus regarding the most preferred to least preferred PC for the proposed computer lab. After the group had come to consensus, respondents were asked to record their own personal judgements. Interestingly enough, most individuals did not change their answers from the group ranking. This could reflect respondent homogeneity, peer-influenced bias, or perhaps lack of dedication to the task.

To review, we hypothesized that the group exercise might better reflect the depth of processing and consideration that respondents would undertake if they were actually making the decision in the real world. It might also mimic the seeking and sharing of information that typically accompanies high involvement purchases. We personally observed most of the Super Holdout sessions, and in our opinion, respondents deliberated with a good deal of effort. However, we cannot know whether we really accomplished our goal of stimulating respondents to make more life-like judgements.

Table 4 shows predictive results for full-profile and ACA.

Table 4

Super Holdout Task Hit Rates

ACA	73.3%
Full-profile	76.7%

n=80

Full-profile does a better job predicting even the Super Holdout Task than ACA, although the margin of victory is slightly less than for predicting the traditional holdout choices. The t-value of the difference in mean prediction is 1.25 (not significant).

Table 5 displays the result which was at the heart of our research: attribute importances for traditional holdout choices versus the Super Holdout Task.

Table 5

Attribute Importances for Holdout Choices

	Standard Choices	**Super Holdout Choices**
Internet Access	25.7	29.0
RAM	23.8	22.9
Price	18.1	16.2
Hard Drive	12.2	7.2
Warranty	8.4	6.5
Processor	9.5	9.4
Lab Assistants	2.3	8.8
TOTAL	**100.0**	**100.0**
STD. DEVIATION	**7.9**	**8.1**

n=80

The precision of estimates is greater for the Standard Choices than the Super Holdout Choices. Recall that each of the 80 respondents completed five standard choice tasks during the computerized survey, but only one Super Holdout Task. Regrettably, the least stable estimates are the most important to our research.

The ratio of the most to least important factor is 11:1 for Standard Choices, and 4:1 for Super Holdout Choices. This conforms to our hypothesis—but focusing only on the extreme points is quite susceptible to error, given the instability in the estimates for the Super Holdout Choices. The standard deviations suggest there is very little difference in the spread of importances between the standard holdouts and the Super Holdout Task.

There are several competing explanations for why we didn't observe significantly flatter results for the Super Holdout Task relative to the traditional holdouts:

1) Maybe people really don't display flatter importances in more carefully considered decisions.

2) Perhaps the students didn't take the Super Holdout Task as seriously as we had hoped.

3) Maybe they did take it seriously, and they also took the standard choice tasks as seriously.

We are more inclined to believe the second explanation. Until more evidence is shown, however, the main hypothesis of our paper remains unproven.

ANATOMY OF ACA IMPORTANCES

While we weren't able to find significant differences in attribute importances between the traditional holdouts and the Super Holdout Task in our pilot study, we confirmed that ACA importances tend to be "flatter" than full-profile importances.

ACA utilities are derived from two sources: the pairs and priors.

> **Priors:** We used default settings for priors, including a 4-point scale for stated importances.
>
> **Pairs:** We again used default settings. Respondents judged pairs on a nine-point graded scale. The design included eighteen total pairs. Twelve pairs were shown on two attributes, and six more pairs on three attributes.

Table 6 displays the importances for ACA as given earlier in Table 2, along with importances derived from just the priors. As before, the importances were computed from average utilities on attributes with assumed *a priori* order.

Table 6

ACA Importances by Component

	ACA Final Importances	**ACA Priors**
Internet Access	21.5	17.4
RAM	19.8	16.6
Price	15.0	15.5
Hard Drive	14.0	14.6
Warranty	11.6	13.0
Processor	10.7	12.9
Lab Assistants	7.4	10.0
TOTAL	**100.0**	**100.0**
STD. DEVIATION	**4.6**	**2.3**

n=80

The importances are less extreme in the priors than in the final utilities. Indeed, with a 4-point importance scale, even if all respondents were in agreement about the most and least important

attributes, the maximum ratio between priors importances would be 4:1.

In Version 4 of ACA, optimal weights for the two components (priors and pairs) are fit to best predict purchase likelihood judgments in the calibration concepts, which for our study were customized to include the six most important attributes for each respondent. For this data set, we may infer that the pairs information is more extreme than both the priors and final optimally-weighted importances.

Critics of ACA have suggested that the 4-point stated importance scale is too coarse. In 1991, Bill McLauchlan reported results from an experiment, which tested different scales for the stated importances for ACA priors (McLauchlan 1991). For that study, the 4-point implementation performed as well (predicting holdout concepts) as customized ACA versions using a 9-point scale and an analog version, which accommodated up to 100 scale points. McLauchlan did not report importances for the attributes involved in his research, however, so we do not know if his study featured such extreme importances as our study.

It seems plausible that designs with greater extremes from the most important to least important attributes might benefit from more than four scale points in the priors section, but until further research is presented on this, it remains speculation. ACA permits the user to customize the scales used in the priors, so one could easily experiment in this area.

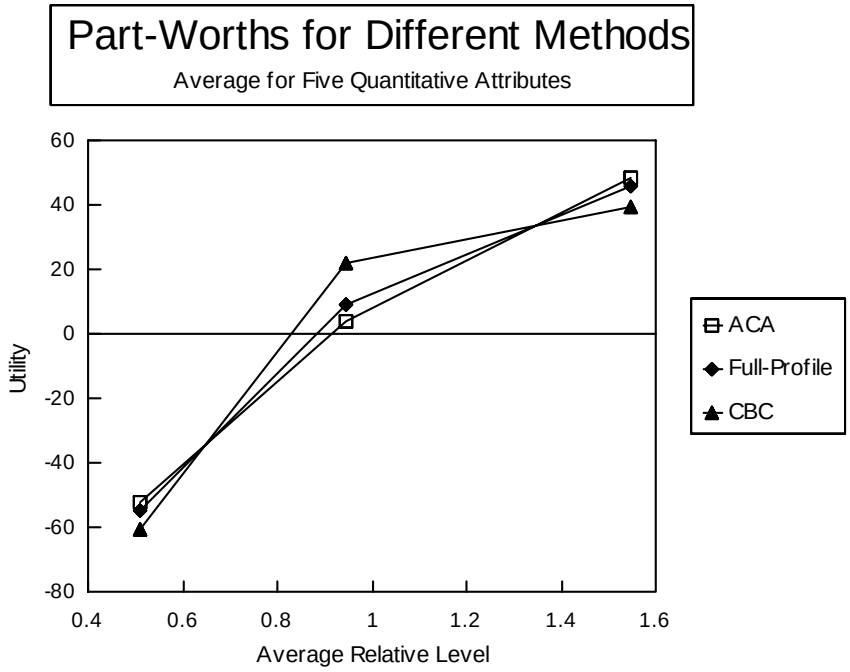
We should again emphasize that this was a pilot study with a small sample and atypical respondents. Lacking evidence about the *true* impact of these attributes on actual purchase decisions for a given product category, one really can't know whether average priors importances are really less valid than importances reflected in the pairs, or the final optimally-weighted result.

ATTRIBUTE PART-WORTHS

Attribute importances reflect utility differences between the best and worst levels for attributes. But importances ignore interior levels. The part-worths for intermediate levels may differ between conjoint methods, and may also account for differences in predictability of holdout concepts. Most of the attributes in our study included a middle level, and five of these were quantitative in nature.

Huber has recently noted differences in part-worths between Choice and traditional conjoint methods. He has found that CBC part-worths tend to show more curvature than full-profile ratings-based conjoint and known utilities (Huber *et al*. 1997). On his advice, we investigated this issue with our data set. The results are summarized in Figure 1.

Figure 1



The CBC part-worths are based on first choices from the traditional choice tasks administered during the disk-based survey. We don't show part-worths for the Super Holdout Task due to instability in the part-worth estimates. The part-worth utilities are zero-centered and scaled so that the difference between the best and worst levels is 100 points.

Part-worths from ACA and full-profile are very similar, with the ACA utilities showing the least amount of curvature on average. This is not surprising given the linearity assumption from the priors. The CBC part-worths displayed the most curvature for all five attributes. In some cases, the curvature was quite pronounced, and the average across the five attributes reflects this trend.

Why do CBC part-worths appear to display more curvature? Huber suggests that respondents may adopt a simplification heuristic that involves scanning choice sets for products with the worst level on key attributes (Huber *et al.* 1997). Respondents focus more on avoiding products with the least preferred levels rather than choosing products with enough good features to surpass some utility threshold, which biases the worst level downward. Under the scaling procedure for the data in Figure 1 which fixes the spread of the worst to best levels, this causes the difference in utility between the best and middle levels to narrow.

It is interesting to note that Johnson and Orme found a similar pattern of curvature when comparing CBC utilities derived only from second choices compared to first choices (Johnson and Orme, 1996). Perhaps the phenomenon which causes second choices to show more curvature than first choices is related to the process which influences CBC part-worths to display more curvature than with ACA or full-profile ratings-based conjoint. This remains conjecture, however, and we look forward to more

research on these issues. Since we do not know the true shape of the part-worths for our study, we can only note the differences between methods without judging which method best predicts real world events.

ATTRIBUTE ORDER EFFECTS

It is no great secret that order effects occur in survey research whenever we present lists of items. Researchers have also reported strong attribute order effects for full-profile conjoint and choice-based conjoint (Johnson 1991, Chrzan 1994), but in general we don't see much attention paid to this in practice.

For one full-profile data set, Johnson reported that attribute order effects accounted for roughly 16 percent of the total error variance of conjoint predictions (Johnson 1991). Since it is not reasonable to expect that respondents encounter attributes in the real world in the same order as seen in conjoint tasks, it is natural that we should control for order effects when comparing the validity of ACA and full-profile judgements.

Due to its adaptive nature in the partial-profile pairs section and the ability to randomize attribute presentation in the priors, ACA should be immune to order effects. Based on the past evidence, we expected that order effects could impact the remaining three aspects of our design: full-profile conjoint, computer-administered holdout choices, and the Super Holdout Task.

For our small pilot study we did not find significant attribute order or task order effects. Enough compelling evidence exists from other studies to suggest that we probably would have discovered significant effects given more data points.

CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

This was a small pilot study to test an approach for improving holdout data and designing better conjoint validity studies. A number of caveats and limitations come to mind:

1) Sample size was small: only 80 respondents.

2) MBAs are not a very representative sample.

3) The nine attributes tested were described in very succinct statements. The respondents already had a high degree of familiarity with the attributes. In the real world, nine-attribute full-profile studies might not always be so manageable for respondents.

4) Not enough data points were collected to gauge whether importances derived from the Super Holdout Task were significantly different from those of traditional holdout choices.

5) Perhaps the Super Holdout Task we implemented failed to create a significantly different (and more realistic) experience than the traditional holdouts administered during the course of the survey.

We hope to see further research done in this area. We cannot stress enough that the ideal validity study would include actual purchase as the holdout criterion. The conjoint community thirsts for this

type of research to be published. Indeed, we might call this the Holy Grail of conjoint validation research. We encourage individuals who have the resources to conduct and publish a carefully designed research study with actual purchase choice as the validation criterion. However, just *one* well-done study with real world purchase data would still leave unanswered questions. The ideal conjoint method for predicting high involvement purchases such as computers or cars may not be ideal for predicting purchases for beverages or bubble gum.

In the absence of actual purchase choice, better validation exercises can be designed for comparing conjoint methods. We can imagine that Super Holdout Tasks could take on many forms—many of which could be more realistic and effective than that which we implemented in this pilot study. Regardless of form, the spirit of the task is to put respondents in the same frame of mind as the real world event and to more closely match the consideration and depth of processing as would be expended in the actual purchase decision. Along with the general design principles we've reviewed, we hope aspects of the Super Holdout Task will be used in methodological studies in the future.

Appendix A

Conjoint Attribute Levels

---

1) Compaq
2) Dell
3) Zeos

4) 90-day warranty
5) 1-yr warranty
6) 2-yr warranty

7) Pentium 100 MHz
8) Pentium 133 MHz
9) Pentium 166 MHz

10) 1 lab assistant
11) 2 lab assistants

12) Ergonomic keyboard/mouse
13) Standard keyboard/mouse

14) 600 Mbyte hard drive
15) 1.2 Gbyte hard drive
16) 2 Gbyte hard drive

17) 8 Mb RAM
18) 16 Mb RAM
19) 32 Mb RAM

20) NO modem/Internet access
21) Modem and Internet access

22) $1,000*
23) $1,500
24) $2,000

---

* Respondents were told that their university had budgeted $30,000 for purchasing PCs. It was stressed that recommending a $2,000 PC would allow only 15 PCs to be purchased for the lab.

The attribute levels were described exactly the same in the ACA, full-profile card-sort and holdout concepts.

Appendix B

Full-profile Card Sort Design

One of the challenges of full-profile designs is to keep the total number of stimuli to a reasonable number while still capturing enough information for stable individual-level utility estimation. This was particularly important for this study since respondents would complete ACA, full-profile and additional holdout tasks.

In a previous comparison of full-profile and ACA, Agarwal and Green (1989) used 18 cards to measure six attributes having three levels each. We used 22 cards in our design. With 22 cards, the number of cards to parameters ratio of 1.47 (22/15) is roughly equivalent to Agarwal and Green's ratio of 1.50 (18/12).

Sawtooth Software's CVA version 2 iterative designer was used to generate the design (shown below), which has a D-efficiency of 95.2% (Kuhfeld *et al*. 1994).

Card 1= 2 6 8 11 12 14 18 20 23
Card 2= 1 5 9 11 13 15 18 20 24
Card 3= 1 6 9 11 13 15 17 21 24
Card 4= 2 4 7 11 13 15 18 21 23
Card 5= 3 6 9 10 12 15 19 21 23
Card 6= 3 5 8 10 13 16 18 21 24
Card 7= 2 6 8 11 13 16 19 21 22
Card 8= 1 5 7 10 13 14 17 21 23
Card 9= 2 6 7 10 12 15 19 20 24
Card10= 1 5 8 10 12 15 19 20 22
Card11= 1 6 7 10 12 16 18 21 22
Card12= 3 4 9 11 12 14 18 21 22
Card13= 3 5 7 11 13 14 19 20 22
Card14= 3 6 7 11 12 16 17 20 24
Card15= 2 5 9 11 12 16 17 21 22
Card16= 2 4 9 10 13 16 19 20 24
Card17= 1 4 8 10 13 16 17 20 23
Card18= 2 5 8 10 12 14 17 21 24
Card19= 2 6 9 10 13 14 18 20 22
Card20= 1 5 9 11 12 16 19 20 23
Card21= 1 4 8 11 12 14 19 21 24
Card22= 3 4 8 10 12 15 17 20 22

Each concept was printed in hard-copy on 3 1/2" x 5" cards. Below is an example:

Card# 1
How likely would you be to recommend...

Dell

2-yr warranty

Pentium 133 MHz

2 lab assistants

Ergonomic keyboard/mouse

600 Mbyte hard drive

16 Mbyte RAM

NO modem/Internet access

$1,500

Write a number below between 0 and 100, where 0 = Definitely would NOT; 100 = definitely WOULD.

Your answer _____

Instructions on the survey disk asked respondents to sort the cards into two piles based on preference, and then to divide those two once again. After sorting the cards into four piles, respondents were instructed to write their evaluations on the cards. Then, the cards were shown on the computer screen one at a time, and respondents were asked to type the answers they wrote for the cards. Respondents were encouraged to modify their answers if they desired as they recorded them.

Respondents were randomly given a set of either blue or yellow hard-copy cards for the full-profile task. The attribute rotations in the two versions were as follows:

Blue	Yellow
A) Brand	A) Brand
B) Warranty	E) Ergonomic Features
C) Processor	F) Hard Drive
D) Lab Assistants	G) RAM
E) Ergonomic Features	H) Internet Access
F) Hard Drive	B) Warranty
G) RAM	C) Processor
H) Internet Access	D) Lab Assistants
I) Price	I) Price

References

Agarwal, Manoj K. and Paul E. Green (1989), "Adaptive Conjoint Analysis Versus Self-Explicated Models: Some Empirical Results," *International Journal of Research in Marketing*.

Chrzan, Keith (1994), "Three Kinds of Order Effects in Choice-Based Conjoint Analysis," *Marketing Letters,* 5:2, April, 165-72.

Huber, Joel, Dick R. Wittink, Richard Johnson, and Richard Miller (1992), "Learning Effects in Preference Tasks: Choice-Based Versus Standard Conjoint," Sawtooth Software Conference Proceedings, 275-82.

Huber, Joel, Dan Ariely, and Gregory Fischer (1997), "The Ability of People to Express Values with Choices, Matching and Ratings," Working Paper, Fuqua School of Business, Duke University.

Johnson, Richard M. (1989), "Assessing the Validity of Conjoint Analysis," *Sawtooth Software Conference Proceedings*, 273-80.

Johnson, Richard M. and Bryan K. Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint?" ART Forum, Beaver Creek, Colorado, June.

Kufeld, Warren, Randall D. Tobias and Mark Garratt (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, (November), 545-57.

McLauchlan, William G. (1991), "Scaling Prior utilities in Sawtooth Software's Adaptive Conjoint Analysis," *Sawtooth Software Conference Proceedings*, 251-68.

Pinnell, Jonathan (1994), "Multistage Conjoint Methods to Measure Price Sensitivity," ART Forum, Beaver Creek, Colorado, June.