



Sawtooth Software

RESEARCH PAPER SERIES

Scaling Prior Utilities in Sawtooth Software's Adaptive Conjoint Analysis

William G. McLauchlan,
McLauchlan & Associates, Inc.
1991

SCALING PRIOR UTILITIES IN SAWTOOTH SOFTWARE'S ADAPTIVE CONJOINT ANALYSIS

William G. McLauchlan
McLauchlan & Associates, Inc.

INTRODUCTION

Background

A variety of research methodologies and data collection techniques are currently employed in commercial and academic applications of conjoint analysis. Included are full-profile, self-explicated, "hybrid," and graded paired-comparison designs. Regardless of the methodology, the ultimate value of any conjoint design resides in its ability to correctly predict choice behavior from the resultant utility data.

Because marketplace choice behavior is always impacted by variables extraneous to the environment in which predictor data are collected, the validity of conjoint techniques is often assessed using internal criteria, such as holdout concepts. The better able we are to predict internal holdout behavior, the more comfortable we feel about a given technique as a basis for making external decisions.

Since its introduction, Sawtooth Software's Adaptive Conjoint Analysis (ACA) has enjoyed a large and loyal following. As such, it is not surprising that ACA has begun to receive attention in the literature and at conferences in terms of its ability to correctly predict choice behavior (Finkbeiner and Platz 1986, Huber and Hansen 1986, Finkbeiner 1988, Herman 1988, Agarwal and Green 1989).

Most recently, a paper in press by Green, Krieger and Agarwal (1990) challenges ACA's utility estimation procedures in a number of areas and offers several suggestions for improving those procedures. One specific criticism of ACA is that attribute importance ratings are "too coarse; only four response values are permitted." (In ACA Version 3.0, prior utilities are calculated by weighting reflected attribute level preference ranks by the importance ratings. The priors are updated in realtime by OLS regression in the graded-pairs section of the interview.) The hypothesized implication is that scale incompatibility between the preference section (ranks), the importance section (4-point ratings), and the graded pairs section (9-point ratings) leads to more extreme response patterns than would be predicted given the objective of ACA to present pairs which are nearly equal in utility. Green *et al* encourage the use of "finer-grained" importance scales (e.g., 1-10 or even analog measurement) as a way of increasing the predictive accuracy of ACA.

The purpose of the study reported here was to empirically evaluate the impact of alternative scaling on the ability of ACA to predict holdout choice behavior. Specifically, with the cooperation of Sawtooth Software, three versions of ACA were obtained and evaluated:

- Current ACA Version 3.0

In this version of ACA, attribute-level preferences are obtained through ranking. Attribute importance is collected using a 4-point rating scale. In the paired tradeoffs section, evaluations are made using a 9-point scale (1= Strongly Prefer Left, 5=No Preference, 9=Strongly Prefer Right). (See Johnson 1988 for a complete description of utility calculations)

- ACA Version 3.0 with a 9-point importance scale

This version is identical to current ACA with the exception that the importance rating is obtained using a 9-point scale (each importance rating is multiplied by 4/9 before the priors are calculated).

- Analog ACA

Prior utilities are elicited directly in this version of ACA. The preferred level within each attribute is assigned a value of 100. Using the cursor keys, the remaining levels are moved on the screen and placed by the respondent along a preference continuum, which ranges from 100 to 0 points (the priors are then rescaled to range from -50 to +50). As such, importance ratings, *per se*, are not evoked.

The subject of the study was consumer preference for grocery store features. A total of 10 attributes were investigated. The attributes and levels were the same regardless of the ACA version (see Table 1).

TABLE 1

Grocery Store Attributes and Levels

BAGGING

You bag your own groceries
Store bags groceries for you

STORE HOURS

Open 24 hours everyday
Open 7 AM to Midnight everyday
Open 7 AM to 9 PM everyday

DRIVE TIME

Less than 5 minutes driving time
5 to 9 minutes driving time
10 to 14 minutes driving time
15 to 29 minutes driving time
30 minutes or more driving time

DOUBLE COUPONS

Double coupons everyday
Double coupons once a week
Does not offer double coupons

FLORAL DEPARTMENT

Floral department with delivery
Floral department; no delivery
Does not have floral department

BRAND NAMES

Brand name products only
Brand name 8 store brands only
Brand name, store brands, generics

PHARMACY

Has a pharmacy
Does not have a pharmacy

BANK

Full-service bank inside store
Store has Automatic Teller Mach.
No bank or ATM inside store

LOCATION

Located in an enclosed mall
Located in a shopping center
"Free-standing" location

CANDY IN CHECKOUT LANES

All checkout lanes have candy
Some checkout lanes have candy
No checkout lanes have candy

Methodology

As described below, the sample was divided into three cells. In- person interviewing was used to collect the data. Respondents were intercepted and screened in six geographically dispersed shopping centers (Detroit, Houston, Oklahoma City, Orlando, Santa Fe, and Staten Island). Those individuals who qualified and agreed to participate were brought back to an enclosed room where the interview was administered on PC. Although an interviewer was present at all times, the respondents keyed-in all answers to the computer-presented questions.

Sample Composition/Size

To participate in the study, respondents had to meet the following qualifications:

- Primary Grocery Shopper
- Between 18 and 65 Years of Age
- Not Competitively Employed
- No Past 3 Months Research Participation

A total of 604 interviews were completed with respondents meeting these qualifications. Interviewing was conducted from October 16 through October 30, 1990.

Design

A three-cell design was used where the cells were differentiated based on the version of ACA to be administered. In each cell, both preceding and subsequent to ACA, each respondent also provided likelihood-to-shop ratings for five grocery store holdout concepts. As indicated in the questionnaire section below, the set of holdout concepts Pre- and Post-ACA were identical. Further, the holdout concepts were the same regardless of the respondent's cell assignment.

The key differences between the ACA versions tested, as well as the number of completed interviews by cell, are as follows:

	Cell 1	Cell 2	Cell 3
Importance Scale	4-Point	9-Point	
Preference Scale	Ranks	Ranks	Analog
Paired-Comparisons	9-Point	9-Point	Analog
Completed Interviews	206	201	197

To facilitate communication of results, Cell 1 will be subsequently referred to as the 4-Point Cell, Cell 2 as the 9-Point Cell, and Cell 3 as the Analog Cell.

Questionnaire

Regardless of cell assignment, the questionnaire sequence and content were the same for all respondents:

- Likelihood to shop ratings (0-100 points) for each of five grocery store holdout profiles (Table 2). All respondents in the study rated the same holdout concepts. The holdouts were initially constructed from a random combination of the attributes and levels included in ACA.
- ACA Preference: Ranks or Analog ratings
Importance: 4-point, 9-point, or Analog ratings
Paired Tradeoffs: 9 point or Analog ratings (11 pairs consisting of 2 attributes each were evaluated by each respondent)
- Likelihood to shop ratings (0 -100 points) for the same five grocery store holdout profiles evaluated prior to ACA
- Profile of grocery store shopped most often
- Demographics

TABLE 2

Holdout Profiles

Profile #1

- * You bag your own groceries
- * Open 7 AM to 9 PM everyday
- * 10 to 14 minutes driving time
- * Double coupons once a week
- * Has a floral department but no delivery
- * Carries brand name, store brands, and generics
- * Does not have a pharmacy
- * Store has an Automatic Teller Machine
- * “Free- standing” location
- * None of the checkout lanes have candy

Profile #2

- * You bag your own groceries
- * Open 24 hours everyday
- * Less than 5 minutes driving time
- * Double coupons everyday
- * Does not have a floral department
- * Carries brand name, store brands, and generics
- * Does not have a pharmacy
- * No bank or Automatic Teller Machine inside store
- * Located in an enclosed mall
- * All of the checkout lanes have candy

Profile #3

- * You bag your own groceries
- * Open 7 AM to Midnight everyday
- * 5 to 9 minutes driving time
- * Does not offer double coupons
- * Has a floral department that will deliver
- * Carries brand name and store brands but not generics
- * Has a pharmacy
- * Store has an Automatic Teller Machine
- * Located in a shopping center
- * Some of the checkout lanes have candy, some do not

(continued)

TABLE 2 (continued)

Profile #4

- * Store bags groceries for you
- * Open 7 AM to Midnight everyday
- * 15 to 29 minutes driving time
- * Double coupons everyday
- * Has a floral department but no delivery
- * Carries brand name and store brands but not generics
- * Has a pharmacy
- * Full-service bank inside the store
- * "Free-standing" location
- * None of the checkout lanes have candy

Profile #5

- * Store bags groceries for you
- * Open 24 hours everyday
- * 10 to 14 minutes driving time
- * Does not offer double coupons
- * Has a floral department but no delivery
- * Carries brand name products but no store brands or generics
- * Has a pharmacy
- * No bank or Automatic Teller Machine inside store
- * Located in an enclosed mall
- * All of the checkout lanes have candy

RESULTS

The results of this study are presented in three sections: Pre-Post Holdout Evaluations, ACA First Choice Results, and Interview Length.

Pre-Post Holdout Evaluations

The validity of any conjoint technique, as reflected by the successful prediction of holdout choice behavior, is mathematically constrained by the reliability of the holdout task. As such, the discussion of the impact of priors scaling on holdout prediction is ultimately based on results among respondents for whom test-retest reliability was high. In the present context, reliability was assessed in three interrelated ways: consistency of preference in the Pre-Post holdout task, Pre-Post product-moment correlations, and an analysis of implied paired-comparisons. This section presents the results of these reliability analyses.

Pre-Post Consistency. For each respondent, the Pre-ACA likelihood-to-shop ratings on the holdout concepts were compared to the Post-ACA ratings on the same set of concepts. Shown below is the proportion of respondents in each cell that gave the highest rating to the same

concept both Pre- and Post-ACA. (In all cells, respondents who had tied first-choice ratings had their first-choice "vote" divided evenly among the tied concepts.) As can be seen, the proportion observed among respondents in the Analog Cell is significantly lower than the proportions of consistent respondents in the 4-Point and 9-Point Cells.

<u>Cell</u>	<u>Pre-Post First Choice Hits (%)</u>
4-Point	65.5
9-Point	60.7
Analog	49.7*

* Significantly lower than the 4-Point and 9-Point results ($p < .05$)

The reliability of the holdout task and the criterion-based validity of ACA will ultimately be impacted by the degree of variation in appeal for the holdout concepts. Better predictive validity should be expected of ACA when the distribution of preference for the holdout concepts is skewed compared to when that distribution is uniform. To provide context for the subsequent discussion of validity, shown below, by cell, are the first-choice distributions for the holdout concepts, both Pre- and Post-ACA.

First Choice Shares of Preference

<u>Concept</u>	<u>4-Point</u>		<u>9-Point</u>		<u>Analog</u>	
	<u>Pre-ACA (%)</u>	<u>Post-ACA (%)</u>	<u>Pre-ACA (%)</u>	<u>Post-ACA (%)</u>	<u>Pre-ACA (%)</u>	<u>Post-ACA (%)</u>
1	5.3	6.8	7.0	9.0	7.1	9.1
2	24.8	25.7	20.9	26.9	23.9	20.8
3	14.6	14.6	17.9	13.9	10.2	19.3
4	38.8	31.6	31.3	29.4	35.5	36.5
5	16.5	21.4	22.9	20.9	23.4	14.2

Pre-Post Product-Moment Correlations. Pearson product-moment correlations were computed for each respondent using the Pre- and Post-ACA likelihood-to-shop ratings. Mean correlations were then calculated by converting the individual Pearson r's to Fisher's Z' scores, averaging over all respondents in a given cell and then converting back to a mean Pearson r. Shown below are the mean Pre-Post correlations by cell. Like the Pre-Post First Choice results discussed above, respondents in the 4-Point and 9-Point Cells were significantly more reliable in their holdout concept ratings than were their counterparts in the Analog Cell.

<u>Cell</u>	<u>Average Pre-Post Correlation</u>
4-Point	.52
9-Point	.52
Analog	.36*

*Significantly lower than the 4-Point and 9-Point results ($p < .05$)

Implied Paired-Comparisons. As a final measure of holdout task reliability, implied paired-comparisons were constructed and tested using a variation on a technique described by Huber and Hansen (1986). In this analysis, each holdout task is viewed as implying 10 paired-comparisons (the first-choice concept is preferred over the remaining four, the second-choice is preferred over the remaining three, and so forth). To assess reliability, 5x5 one-zero (preferred-not preferred) paired comparison matrices were composed for each holdout task for each respondent. The matrices were then multiplied using Boolean logic to produce individual respondent paired-comparison "hits" matrices. For each respondent, then, the proportion of correct Pre-Post implied paired-comparisons is calculated.

As an example of the calculations, if a respondent's holdout ratings were:

	Concept				
	1	2	3	4	5
Pre-ACA	20	15	50	100	90
Post-ACA	25	30	65	90	95

then the implied paired-comparison matrices would be:

	Pre-ACA Concept						Post-ACA Concept				
	1	2	3	4	5		1	2	3	4	5
	<hr/>						<hr/>				
1	-	0	1	1	1	1	-	1	1	1	1
2	1	-	1	1	1	2	0	-	1	1	1
3	0	0	-	1	1	3	0	0	-	1	1
4	0	0	0	-	0	4	0	0	0	-	1
5	0	0	0	1	-	5	0	0	0	0	-

and the “hits“ matrix would be:

	Concept				
	1	2	3	4	5
	<hr/>				
1	-	0	1	1	1
2	0	-	1	1	1
3	1	1	-	1	1
4	1	1	1	-	0
5	1	1	1	0	-

The proportion of implied paired-comparison “hits“ would be $16/20 = 80\%$.

The observed average proportions by cell are shown below and are not significantly different at the $p=.05$ level.

<u>Cell</u>	<u>Average Correct Implied Paired-Comparisons</u> (%)
4-Point	65.8
9-Point	64.9
Analog	60.5

ACA First Choice Results

The ability of ACA to correctly predict holdout choice behavior was assessed in three ways: first choice hits predicted by ACA, holdout task variance explained by ACA, and an estimation of the proportion of error variance in ACA due to the unreliability of the holdout task.

First Choice Hits. For each respondent, the ACA-predicted first-choice profile was compared with the highest rated profile in each of the two holdout tasks. Shown below are the proportions of respondents in each cell for whom ACA correctly predicted holdout first choices (ties are counted as "hits" throughout). There are no significant differences among cells in the first choice accuracy of ACA for either holdout task.

<u>Cell</u>	<u>Proportion of First Choice ACA Hits</u>	
	<u>Pre-ACA</u> <u>Holdout</u> (%)	<u>Post-ACA</u> <u>Holdout</u> (%)
4-Point (n=206)	34.5	36.9
9-Point (n=201)	35.3	37.8
Analog (n=197)	33.5	32.5

As previously noted, respondents in the 4-Point and 9-Point Cells were significantly more reliable in the Pre-Post ACA holdout tasks than were the respondents in the Analog Cell. However, when the data for consistent respondents are isolated and the holdout first choices are compared with ACA-predicted first choices, the resultant proportions for each cell are identical.

<u>Cell</u>	<u>Proportion of First Choice Hits</u> <u>Among Consistent Respondents</u> (%)
4-Point (n=135)	45.9
9-Point (n=122)	45.9
Analog (n=98)	45.9

Variance Explained. Mean r-squared values were computed between the holdout task likelihood-to-shop ratings and the corresponding sums of the ACA partworths. The analysis was conducted among the total sample and among those respondents consistent in their Pre-Post choices. In all instances, r-squares are based on product-moment correlations corrected for attenuation. Shown in Table 3 are the results of these analyses. None of the between, or within, cell differences are significant at p=.05 level.

TABLE 3
Variance Explained in the Holdout Tasks by ACA

<u>Total Sample Average R-Squared</u>				
<u>Cell</u>	<u>ACA and Pre-ACA Holdout Task</u>	<u>ACA and Post-ACA Holdout Task</u>	<u>Post-Pre Difference</u>	<u>Mean</u>
4-Point (n=206)	.64	.71	+.07	.68
9-Point (n=201)	.64	.64	.00	.64
Analog (n=197)	.67	.69	+.02	.68
<u>Pre-Post Holdout Consistent Average R-Squared</u>				
<u>Cell</u>	<u>ACA and Pre-ACA Holdout Task</u>	<u>ACA and Post-ACA Holdout Task</u>	<u>Post-Pre Difference</u>	<u>Mean</u>
4-Point (n=126)	.63	.69	+.06	.66
9-Point (n=112)	.56	.59	+.03	.58
Analog (n=88)	.65	.68	+.03	.66

To provide insight into how great the impact of holdout task unreliability can be on the estimates of the holdout variance explained by ACA, the observed mean R-squared values uncorrected for attenuation are reported below.

<u>Cell</u>	<u>Total Sample</u>	<u>Holdout Consistent</u>
4-Point	.18	.23
9-Point	.08	.16
Analog	.16	.27

The predictive improvement in holdout behavior, over chance, that can be attributed to ACA can be inferred from $1 - \sqrt{(1 - r^2)}$. This calculation reveals the extent to which error in criterion prediction is reduced by ACA. When $1 - \sqrt{(1 - r^2)}$ equals 0, the error in ACA prediction will equal the error observed if holdout preferences were simply guessed. Shown below are the results of this calculation for the total sample and for respondents who were consistent in their first choice holdout behavior.

	<u>Total Sample</u>	<u>Holdout Consistents</u>
	$1 - \sqrt{(1 - r^2)}$	$1 - \sqrt{(1 - r^2)}$
4-Point	.43	.42
9-Point	.40	.35
Analog	.43	.42

Estimating ACA Error Variance. As an extension to the above analyses, the relationships between ACA predictions and holdout task behavior were evaluated using an approach described by Johnson (1989). This analysis consisted of the following calculations:

1. Estimation of the correlation between the ACA predictions and a perfectly reliable choice measure.
2. Estimation of the error variance in ACA prediction that would be expected if the holdout tasks were perfectly reliable.
3. Calculation of the observed ACA error variance (uncorrected for attenuation).
4. Calculation of the proportion of observed ACA error variance due to unreliability of the holdout tasks.

Results of these calculations are shown in Table 4 for both the total sample and for those respondents who were first-choice consistent at the holdout task. As can be seen, for the total sample, and also for those respondents who were holdout consistent, more than half of the error observed in ACA's predictions is apparently due to unreliability in ratings of the holdout concepts.

Table 4

Estimation of Error Variance in ACA

Total Sample

	Cell		
	<u>4-Point</u> (206)	<u>9-Point</u> (201)	<u>Analog</u> (197)
Estimated r Between ACA and Perfectly Reliable Choice Measure	.82	.80	.82
Relative Error Variance in ACA if Holdout were Perfectly Reliable	.32	.36	.32
Observed Error Variance in ACA	.82	.92*	.84
Proportion of ACA Error Variance due to Unreliability in Holdout Task	.61	.61	.62

* Significantly higher than the 4-Point and Analog results ($p < .05$)

Pre-Post Holdout Consistent Respondents

	Cell		
	<u>4-Point</u> (126)	<u>9-Point</u> (112)	<u>Analog</u> (88)
Estimated r Between ACA and Perfectly Reliable Choice Measure	.81	.76	.82
Relative Error Variance in ACA if Holdout were Perfectly Reliable	.34	.42	.34
Observed Error Variance in ACA	.77	.84	.73
Proportion of ACA Error Variance due to Unreliability in Holdout Task	.56	.50	.54

Interview length

To determine if interview length would vary as a function of the type of scaling, the time spent completing the ACA portion of the interview was measured for each respondent using the PC's internal clock. Shown below, by cell, are the mean times in minutes that the ACA portion of the interview took to complete.

<u>Cell</u>	<u>Average Length of ACA Interview</u>
4-Point	7.2 minutes
9-Point	7.4 minutes
Analog	11.5 minutes*

*Significantly higher than the 4-Point and Analog results (p<.05)

DISCUSSION

Validity

The results of this study refute the contention that scaling incompatibilities lead to degradation in the ability of ACA to correctly predict holdout behavior. This conclusion is based on the following key findings:

- There are no significant differences among cells in the proportions of first-choice hits by ACA for either the Pre-ACA or Post-ACA holdout tasks.
- The variance in the holdout ratings explained by ACA is constant across the three cells.
- The prediction error attributable to ACA does not vary as a function of scaling differences.
- The estimated error variances in ACA, given a perfectly reliable holdout task, are equal regardless of the type of scaling.

In spite of the results of this study, it might still be argued that constancy in scaling, if nothing else, would be less disconcerting to a respondent as he/she shifts from question type to question type as the interview proceeds. If the psychological comfort of the respondent is of concern, the point would be well taken. However, there is nothing in the data reported here to suggest that ACA suffers because of variations in scaling.

As noted above, shifting to the "finest" level of scaling significantly increases the length of the ACA portion of the interview. In the absence of evidence to support the hypothesis that finer scaling would improve the validity of ACA, the longer interview length is, by itself, sufficient reason to avoid changing ACA scales to analog measurement.

Reliability

As shown by the following findings, the reliability of the holdout task degrades as ACA scaling becomes "finer."

- The proportion of Pre-Post holdout hits in the Analog Cell was significantly lower than the proportions observed in the 4-Point and 9-Point Cells.
- The Analog Cell Pre-Post holdout ratings were significantly less correlated than were the ratings in both the 4-Point and 9-Point Cells.

It could be hypothesized, in the same spirit that motivated this research, that the significantly lower holdout reliability in the Analog cell results from the psychological, if not mathematical, impact of variations in scaling between the holdout tasks and the ACA interview. The reliability may have been higher if the holdout ratings had been collected using analog measurement as well.

Because predictive validity is constrained by the reliability of the criterion variables, efforts to improve ACA (or any conjoint technique) are best assessed in an environment where the criterion is measured as reliably as possible. Anastasi (1976) has stated that the criteria most often used in validating tests are, in themselves, dynamic rather than static. As a consequence, criterion-based validity will always be transient. Furthermore, Johnson (1990) reported that the ACA experience can actually help respondents clarify their value systems.

The ACA interview should be expected, then, to have a bearing on the test-retest reliability of the holdout task. As such, the Post-ACA holdout tasks should be better predicted by ACA than are Pre-ACA tasks. In fact, the variance explained in the Post-ACA holdout ratings by the ACA predictions is higher, though not significantly so, in all three cells among consistent respondents.

Finally, ACA is too often held up against paper-and-pencil conjoint techniques (and vice versa) in validity “showdowns” (Herman, 1990). We may be in danger of losing sight of a more fundamental challenge: developing a better understanding of the mitigating effects of the techniques, themselves, on individual value systems and therefore on results. Do respondents “learn” in full-profile ranking or rating tasks? Probably not. Are value systems impacted by the ACA experience? Probably so. Definitive answers to these questions will evolve only from further empirical research and not from hyperbole.

REFERENCES

- Agarwal, M. K. & Green, P. E. "Adaptive Conjoint Analysis versus self-explicated models: Some empirical results," Working paper, State University of New York at Binghamton, 1989.
- Anastasi, A. *Psychological testing* (4th ed.). New York: Macmillan, 1976.
- Finkbeiner, C. "Comparison of conjoint choice simulators," In *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, 1988.
- Finkbeiner, C. & Platz, P. "Computerized versus paper and pencil methods: A comparison study," Paper presented at the Association for Consumer Research Conference, Toronto, 1986.
- Green, P. E., Krieger, A. M., & Agarwal, M. K. "Adaptive Conjoint Analysis: Some caveats and suggestions," *Journal of Marketing Research* (in press).
- Herman, S. "Software for full-profile conjoint analysis." In *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, 1988.
- Herman, S. "Notes on conjoint analysis," Issue 2. New York: Bretton-Clark, 1990.
- Huber, J. & Hansen, D. "Testing the impact of dimensional complexity and affective differences on paired concepts in Adaptive Conjoint Analysis." In M. Wallendorf & P. Anderson, P. (Eds.), *Advances in Consumer Research*. Provo, UT: Association for Consumer Research, 1986.
- Johnson, R. M. Comment on Green *et al.* *Journal of Marketing Research* (in press).
- Johnson, R. M. "Assessing the Validity of Conjoint Analysis." In *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, 1989.
- Johnson, R. M. "Adaptive Conjoint Analysis," In *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, 1987.